




3 1761 10374385 2









Digitized by the Internet Archive  
in 2023 with funding from  
University of Toronto

<https://archive.org/details/31761103743852>





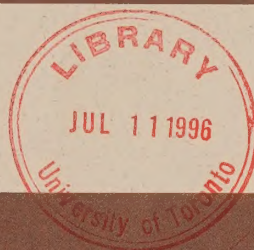


12  
-001



107  
Government  
Publications

# SURVEY METHODOLOGY



Catalogue 12-001

A JOURNAL  
PUBLISHED BY  
STATISTICS CANADA

JUNE 1996

•

VOLUME 22

•

NUMBER 1



Statistics  
Canada

Statistique  
Canada

Canada









---

# SURVEY METHODOLOGY

---

A JOURNAL  
PUBLISHED BY  
STATISTICS CANADA

JUNE 1996 • VOLUME 22 • NUMBER 1

Published by authority of the Minister  
responsible for Statistics Canada

© Minister of Industry, 1996

All rights reserved. No part of this publication may be reproduced,  
stored in a retrieval system or transmitted in any form or by any  
means, electronic, mechanical, photocopying, recording or otherwise  
without prior written permission from Licence Services,  
Marketing Division, Statistics Canada,  
Ottawa, Ontario, Canada K1A 0T6.

June 1996

Price: Canada: \$45.00

United States: US\$50.00

Other countries: US\$55.00

Catalogue no. 12-001-XPB

Frequency: Semi-annual

ISSN 0714-0045

Ottawa



Statistics  
Canada

Statistique  
Canada

Canada



# SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is abstracted in The Survey Statistician and Statistical Theory and Methods Abstracts and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

## MANAGEMENT BOARD

**Chairman** G.J. Brackstone

**Members** D. Binder R. Platek (Past Chairman)  
G.J.C. Hole D. Roy  
F. Mayda (Production Manager) M.P. Singh  
C. Patrick

## EDITORIAL BOARD

**Editor** M.P. Singh, *Statistics Canada*

### Associate Editors

D.R. Bellhouse, *University of Western Ontario*  
D. Binder, *Statistics Canada*  
J.-C. Deville, *INSEE*  
J.D. Drew, *Statistics Canada*  
J.-J. Droesbeke, *Université Libre de Bruxelles*  
W.A. Fuller, *Iowa State University*  
M. Gonzalez, *U.S. Office of Management and Budget*  
R.M. Groves, *University of Maryland*  
M.A. Hidirolou, *Statistics Canada*  
D. Holt, *Central Statistical Office, U.K.*  
G. Kalton, *Westat, Inc.*  
A. Mason, *East-West Center*  
D. Pfeffermann, *Hebrew University*

J.N.K. Rao, *Carleton University*  
L.-P. Rivest, *Université Laval*  
I. Sande, *Bell Communications Research, U.S.A.*  
C.-E. Särndal, *Université de Montréal*  
W.L. Schaible, *U.S. Bureau of Labor Statistics*  
F.J. Scheuren, *George Washington University*  
J. Sedransk, *State University of New York*  
C.J. Skinner, *University of Southampton*  
P.J. Waite, *U.S. Bureau of the Census*  
J. Waksberg, *Westat, Inc.*  
K.M. Wolter, *National Opinion Research Center*  
A. Zaslavsky, *Harvard University*

**Assistant Editors** J. Denis, M. Latouche, H. Mantel and D. Stukel, *Statistics Canada*

---

## EDITORIAL POLICY

Survey Methodology publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

### Submission of Manuscripts

Survey Methodology is published twice a year. Authors are invited to submit their manuscripts in either English or French to the Editor, Dr. M.P. Singh, Household Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Four nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

### Subscription Rates

The price of Survey Methodology (Catalogue no. 12-001-XPB) is \$45 per year in Canada, US \$50 in the United States, and US \$55 per year for other countries. Subscription order should be sent to Statistics Canada, Operations and Integration Division, Circulation Management, 120 Parkdale Avenue, Ottawa, Ontario, Canada K1A 0T6. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, and the Statistical Society of Canada.



# **SURVEY METHODOLOGY**

**A Journal Published by Statistics Canada**

**Volume 22, Number 1, June 1996**

## **CONTENTS**

In This Issue .....	1
D.G. STEEL, D. HOLT and M. TRANMER Making Unit-Level Inferences From Aggregated Data .....	3
D.A. BINDER Linearization Methods for Single Phase and Two-Phase Samples: A Cookbook Approach .....	17
W. YUNG and J.N.K. RAO Jackknife Linearization Variance Estimators Under Stratified Multi-Stage Sampling ..	23
Y.P. CHAUBEY, F. NEBEBE and P.S. CHEN Small Area Estimation Under an Inverse Gaussian Model .....	33
L. RIZZO, G. KALTON and J.M. BRICK A Comparison of Some Weighting Adjustment Methods for Panel Nonresponse .....	43
Y. DING and S.E. FIENBERG Multiple Sample Estimation of Population and Census Undercount in the Presence of Matching Errors .....	55
J.G. SLANTA and T.R. KRENZKE Applying the Lavallée and Hidioglou Method to Obtain Stratification Boundaries for the Census Bureau's Annual Capital Expenditures Survey .....	65
E.B. DAGUM A New Method to Reduce Unwanted Ripples and Revisions in Trend-Cycle Estimates From X-11-ARIMA .....	77
Y. TILLÉ A Moving Stratification Algorithm .....	85
A.G. de WAAL and L.C.R.J. WILLENBORG A View on Statistical Disclosure Control for Microdata .....	95







## In This Issue

This issue of *Survey Methodology* contains articles dealing with a variety of subjects. In the first article, Steel, Holt and Tranmer examine the problem of using aggregated data in studies on relationships at the individual or household level. They propose a simple general model that seeks to take account of the geographical effects of aggregation. They then describe how this model effects both the estimation of population means and covariance matrices and analysis at the regional level. In addition, by introducing auxiliary variables for which certain external sources provide an estimate of the covariance matrix at the unit level, the authors propose methods that provide an unbiased estimate of the parameters at the individual level, so as to avoid the effect of geographical aggregation.

Binder gives a “cookbook” approach for deriving Taylor series approximations to the variances of a wide class of estimators from complex surveys. Several useful examples are presented, as well as new results on the application of this general technique to two-phase sampling. A justification of this method is given, showing the procedure to be consistent with the formulation given in earlier work by Binder and Patak.

Yung and Rao suggest a linear approximation to the jackknife variance estimator. This linearized jackknife inherits the good statistical properties of the usual jackknife variance estimator but is computationally much less intensive. The specific form of the proposed variance estimator is developed for the generalized regression estimator of a total and for the ratio of two generalized regression estimators. In a simulation study using data from the U.S. Current Population Survey, they found that the jackknife, the linearized jackknife, and the usual linearization variance estimators worked quite well for poststratified estimates of a total, while an incorrect form of the jackknife was badly biased.

Chaubey, Nebebe and Chen consider use of an Inverse Gaussian model for positively skewed data and develop a corresponding model assisted estimators for domain totals, which consist of Inverse Gaussian regression predictors together with an expansion estimators of the regression bias. A modified version of the estimator which gives reduced weight to the bias correction term, analogous to a modified regression estimator proposed by Särndal and Hidiroglou, is also proposed. In a simulation study using synthetic income data based on Statistics Canada’s Survey of Household Income, Facilities and Finance the proposed estimators are found to work reasonably well.

Rizzo, Kalton and Brick investigate the use of auxiliary information in compensating for panel nonresponse through weight adjustment techniques. Using data from the Survey of Income and Program Participation (SIPP) to illustrate, they address two important issues, namely, the choice of auxiliary variables to be used in a nonresponse weight adjustment technique, and the choice of technique itself. A screening procedure in conjunction with logistic regression modelling are the means by which appropriate auxiliary variables are chosen. The nonresponse weighting adjustment methods considered are based on logistic regression models, categorical search algorithms and generalized raking. An empirical comparison of the various methods is discussed in detail.

Ding and Fienberg develop models of matching error which can be used in estimation of total population from a probabilistic match of two or more samples. They develop their models for the particular application of a multiple sample census, that is, a census supplemented by auxiliary samples. They illustrate the usefulness of their methods by applying them in an analysis of the 1988 St. Louis Dress Rehearsal Census data for which three samples were matched: the Census itself, the Post Enumeration Survey sample, and the Administrative List Supplement.

In a paper on optimal stratification, Slanta and Krenzke talk about the use of the Lavallée-Hidiroglou method. This iterative method minimizes the sample size while fixing the coefficient of variation. In a practical illustration, the authors present the difficulties with the Lavallée-Hidiroglou method and show how they were resolved.

Dagum proposes a new method for estimating underlying trends from seasonally adjusted data. The approach consists of two steps. The seasonally adjusted data are first extrapolated based on an ARIMA model. A 13-term Henderson filter is then applied to the extended series, using strict sigma limits for the identification and replacement of extreme values. The new method is compared to the standard method using data from several economic time series. It is found that the new method produces fewer unwanted ripples in the estimated trend, while identifying turning points as just quickly and requiring smaller revisions on average.

Tillé proposes an algorithm that generalizes the selection-rejection method used for constructing a simple random sample without replacement. A specific case of this algorithm, which is called the “mobile stratification algorithm”, is discussed. It serves to obtain a smoothed stratification effect by using as a stratification variable the serial number of the units of observation. This algorithm gets around the thorny problem of a continuous variable in strata.

De Waal and Willenborg review recent research on statistical disclosure control for microdata files from the perspective of Statistics Netherlands. Models are developed for the probability that a particular record could be re-identified and for the probability that some record in a microdata file could be re-identified. Global recoding and local suppression are considered as methods to reduce disclosure risk. They conclude that there is still much need for further methodological research and development of efficient software.

Finally, it is with sadness that I note the recent passing away of Maria Gonzalez, who died of cardiac arrest while vacationing in Puerto Rico this past February. Among her many contributions to the statistical community, for the past several years Maria has been an Associate Editor for the *Survey Methodology* journal. Her contribution in this capacity to the quality and breadth of this journal was very much appreciated, and she will be sorely missed. An obituary, written by Elizabeth and Fritz Scheuren, appeared in the April issue of *Amstat News*.

The Editor



# Making Unit-Level Inferences From Aggregated Data

D.G. STEEL, D. HOLT and M. TRANMER<sup>1</sup>

## ABSTRACT

Data are often available only as a set of group or area means. However, it is well known that statistical analysis based on such data will often produce results very different from those obtained from analysing the corresponding individual or household data. If the results of area level analyses are thought to apply to the individual level then we risk committing the ecological fallacy. Aggregation or ecological effects arise in part because geographic areas are not comprised of random groupings of people or households but exhibit strong socio-economic differences between areas. The population structure must be incorporated into the statistical model underpinning the analysis if aggregation effects are to be understood. A simple general model is proposed to achieve this and the consequences of the model and its implications for the estimation of population means and covariance matrices are obtained. Furthermore, methods are suggested which can provide unbiased estimates of individual level parameters from aggregated data and so avoid the ecological fallacy. These methods rely on identifying the “grouping variables” that characterise the process that led to the population structure, or at least characterise the area differences. An estimate of the unit level covariance matrix of the grouping variables is required from some source. Data from the 1991 Census of the United Kingdom have been analysed to identify the important grouping variables and evaluate the effectiveness of the proposed adjustment methods for the estimation of covariance matrices and correlation coefficients. These results lead to a suggested strategy for the analysis of aggregated data.

**KEY WORDS:** Aggregation; Ecological fallacy; Grouping; Selection; Variance components.

## 1. INTRODUCTION

Researchers are often faced with the problem of wishing to investigate individual level relationships but having to make use of aggregated data, such as the means or totals for geographic areas. Ideally unit level data collected in a sample survey or census would be used, but may not be accessible because of confidentiality restrictions, or because the variables have not been collected in a recent survey or census. Administrative systems provide information on a range of variables, for example on unemployment, health, morbidity, but because of confidentiality requirements these data are usually made available for aggregates, such as geographic areas. The census also provides data for geographic areas. For these reasons, analysis of group level data is still an option used widely in social and epidemiological research.

Consider a population in which each individual has associated a vector of variables of interest, whose distribution has mean  $\mu_y$  and covariance matrix  $\Sigma_{yy}$ . We are interested in relationships among the variables of interest as reflected by correlations, regression coefficients and principal components, which may all be derived from the covariance matrix,  $\Sigma_{yy}$ , which is our basic target of inference. For example, the variables of interest might include a set of attainment tests in an educational study; the incidence of a particular disease and a set of explanatory variables in an epidemiological study; or a set of

deprivation measures in a sociological study. We suppose that individual level data are unavailable. However, the region may be subdivided into a set of small areas such as Census Enumeration Districts (EDs), and for each small area,  $g$ , or for a sample of areas, we observe the vector of average values  $\bar{y}_g$  for the variables of interest together with the sample size  $n_g$  on which this is based.

The objective of the analysis,  $\Sigma_{yy}$ , is a covariance matrix which spans the small areas. The target of inference is not conditional on small area membership but refers to the marginal distribution across small areas. This contrasts with situations, such as small area estimation, in which the target of inference is in the conditional distribution given the small area. This is a separate, legitimate objective with which we are not concerned. The same models may be applicable, but the targets of inference are different. However, our formulation does allow for group specific variables to be included as variables of interest if required. For example, if we associate with each individual a set of ED means for the area in which the individual is located, then these can be included within the vector,  $y$ , of interest. In particular, regression analyses which include small area means as explanatory variables in the regression model can be encompassed by the approach.

The literature associated with the analysis of aggregated data dates back to Gehlke and Biehl (1934) and includes significant contributions by Yule and Kendall (1950) and Robinson (1950), Blalock (1964), Openshaw and Taylor

<sup>1</sup> D.G. Steel, Department of Applied Statistics, University of Wollongong, NSW 2522, Australia; D. Holt and M. Tranmer, Department of Social Statistics, University of Southampton, S017 1BJ, United Kingdom.

(1979) and more recently Arbia (1989). There are also problems associated with the fact that the areal units used often have no special significance, being constructed for reasons of cost, operational or administrative convenience. Moreover, the results of the group level analysis will depend on the scale of the units, that is their average size and the particular set of boundaries chosen. Several empirical studies have demonstrated these effects, including Clark and Avery (1976), Perle (1977), Openshaw (1984), and Fotheringham and Wong (1991). However, these studies have not provided any generally applicable theory or practical methods of modifying the results of group level analyses to provide reliable unit level inferences.

Aggregation effects arise because geographic units are not comprised of random groupings of people. Individuals in the same area generally tend to be more alike because they choose to live in areas in a non-random way, or because they are subjected to common influences, or because they interact with one another. Thus there are socio-economic differences between areas which are confounded with the individual effects in any statistical analysis performed using aggregated data for the areas. A simple general model is proposed which seeks to incorporate these effects. The consequences of this model and its implications for area level analysis are obtained. Furthermore, methods are suggested which provide, under certain circumstances, unbiased estimates of individual level parameters from aggregated level data and so avoid the ecological fallacy. These methods involve auxiliary variables for which a unit level sample covariance matrix is available from some source. This approach has been applied to data from the 1991 Census of the United Kingdom and a strategy developed for the analysis of aggregated data.

## 2. MODELS FOR AREA EFFECTS

We consider a population of  $N$  individuals each having a vector  $y$  of characteristics of interest. The population is comprised of  $M$  groups and the random variable  $c_i$  indicates the area to which the  $i$ -th population unit belongs. The number of individuals in the  $g$ -th area is  $N_g$ .

We consider  $\mu_y$  and  $\Sigma_{yy}$  to be superpopulation parameters and the following statistical theory is obtained in this framework. However, we consider some survey design issues at the end of section 2.

We assume that there exists a sample data set  $s$  of size  $n$  and that these individual data have been aggregated to provide a set of  $m$  area means which are available for analysis. The following area level statistics can be calculated:

the  $g$ -th area mean:

$$\bar{y}_g = \frac{1}{n_g} \sum_{i \in g, s} y_i \quad (2.1)$$

the overall sample mean:

$$\bar{y} = \frac{1}{n} \sum_{g \in s} n_g \bar{y}_g = \frac{1}{n} \sum_{i \in s} y_i \quad (2.2)$$

the area level sample covariance matrix:

$$\bar{S}_{yy} = \frac{1}{m-1} \sum_{g \in s} n_g (\bar{y}_g - \bar{y}) (\bar{y}_g - \bar{y})'. \quad (2.3)$$

Analogous unit level statistics may be defined but will be unavailable to the analyst. For example  $S_{yy} = 1/(n-1) \sum_{i \in s} (y_i - \bar{y})(y_i - \bar{y})'$  is the unit level sample covariance matrix.

### 2.1 Random Grouping

While geographic groups are rarely formed randomly, such a situation is a useful starting point in considering ecological analysis. If groups are randomly formed then many group level analyses are valid, albeit with a reduced efficiency. Steel and Holt (1995) consider the properties of statistics such as means, variances, regression and correlation coefficients in this situation. When the groups are randomly formed *i.e.*,  $y \perp c$  then

$$E[\bar{y}_g | s, c] = \mu_y \quad (2.4)$$

$$V(\bar{y}_g | s, c) = \frac{1}{n_g} \Sigma_{yy}. \quad (2.5)$$

The basic properties of the unit and group level statistics then follow readily

$$\text{Cov}(\bar{y}_g, \bar{y}_h | s, c) = 0 \quad g \neq h \quad (2.6)$$

$$E[\bar{y} | s, c] = \mu_y \quad (2.7)$$

$$E[S_{yy} | s, c] = \Sigma_{yy} \quad (2.8)$$

$$E[\bar{S}_{yy} | s, c] = \Sigma_{yy}. \quad (2.9)$$

These properties apply if the sampling is ignorable given the group indicatives, which means the sample design can depend on the groups but not on  $y$  or any variable which is related to  $y$  conditional on  $c$ . For example a census or a simple random sample of groups and units within groups may be used.

Unweighted group level statistics may be used by setting  $n_g = 1$  in equations (2.2) and (2.3). This leads to inefficient estimators. The degree of inefficiency will depend on the distribution of the group sample sizes. Weighting by the group sample sizes is important and when this is done



inference can proceed as usual with appropriate adjustments to the degrees of freedom. Variability is determined by the number of areas rather than the number of individual observations and confidence intervals and tests are adjusted accordingly.

## 2.2 A Variance Component Model

A simple way to represent the positive intra-group correlation that is usually observed in grouped populations is through a variance components model, which in the multivariate case corresponds to

$$y_i = \mu_y + v_g + \epsilon_i \quad i \in g$$

where  $v_g$  and  $\epsilon_i$  are independent random components at the group and individual level respectively, both with zero expectation,  $V(\epsilon_i | c) = \Sigma_{\epsilon\epsilon}$  and  $V(v_g | c) = \Delta_{yy}$ .

**Model A:**

$$E[y_i | c] = \mu_y \quad (2.10)$$

$$V(y_i | c) = \Sigma_{\epsilon\epsilon} + \Delta_{yy} = \Sigma_{yy} \quad (2.11)$$

$$\begin{aligned} \text{Cov}(y_i, y_j | c) &= \Delta_{yy} \quad \text{if } c_i = c_j \quad i \neq j \\ &= 0 \quad \text{otherwise.} \end{aligned} \quad (2.12)$$

The notation  $V(\cdot | c)$  implies the covariance matrix conditional on the group labels  $c$  and hence determines common group membership. It is, however, taken to be unconditional over the group level random effects. Thus  $V(y_i | c)$  contains the total variance from both the within group covariance matrix  $\Sigma_{\epsilon\epsilon}$  and the group level covariance matrix  $\Delta_{yy}$ .

The properties of the sample group level means follow readily from Model A, if the sampling is ignorable given  $c$ ,

$$E[\bar{y}_g | s, c] = \mu_y \quad (2.13)$$

$$V(\bar{y}_g | s, c) = \frac{1}{n_g} (\Sigma_{yy} + (n_g - 1)\Delta_{yy}) \quad (2.14)$$

$$\text{Cov}(\bar{y}_g, \bar{y}_h | s, c) = 0 \quad g \neq h. \quad (2.15)$$

The properties of the unit level and group level statistics are

$$E[\bar{y} | s, c] = \mu_y \quad (2.16)$$

$$E[S_{yy} | s, c] = \Sigma_{yy} - \frac{\bar{n}^0 - 1}{n - 1} \Delta_{yy} \quad (2.17)$$

$$E[\bar{S}_{yy} | s, c] = \Sigma_{yy} + (\bar{n}^* - 1) \Delta_{yy} \quad (2.18)$$

where  $\bar{n} = n/m$ ,  $\bar{n}^0 = 1/n \sum_{g \in s} n_g^2 = \bar{n}(1 + C_n^2)$ ,  $\bar{n}^* = \bar{n}(1 - C_n^2/(m - 1))$  and  $C_n^2 = 1/m \sum_{g \in s} (n_g - \bar{n})^2 / \bar{n}^2$  is the square of the coefficient of variation of the group sample sizes in the sample. We note that the coefficient of  $\Delta_{yy}$  is  $0(m^{-1})$  in (2.17) but is  $0(\bar{n})$  in (2.18). This illustrates how a small bias in the unit level analysis can be magnified into a much larger bias in the aggregate level analysis. We will discuss these results further in section 2.4.

## 2.3 Grouping Models

In the discussion of ecological analysis, models have been proposed which take into account the group formation process. In this approach it is assumed that there is a grouping process which allocates individual units to groups according to a vector of grouping variables,  $z_i$ , either stochastically or deterministically. This approach is implicit in Blalock's (1964) analysis and used explicitly by Hannan and Burstein (1974), Litchman (1974), Langbein and Litchman (1978), Smith (1977) and Blalock (1979, 1985). Steel (1985) refers to these models as grouping models since it is assumed that groups are formed by some process involving the variables in the relationships under study. The grouping is seen as a distorting effect and the relationships of interest are defined before the grouping has occurred. It is often noted in the discussion of contextual models that apparent contextual effects may in fact be due to such factors. The multivariate version of this model is:

**Model B:**

$$E[y_i | z, c] = \mu_{y,z} + \beta'_{yz} z_i \quad (2.19)$$

$$V(y_i | z, c) = \Sigma_{yy,z} \quad (2.20)$$

$$\text{Cov}(y_i, y_j | z, c) = 0 \quad i \neq j. \quad (2.21)$$

In this model the conditional expectation of  $y_i$  depends only on the value of the auxiliary variables for the  $i$ -th unit and is independent of the group to which the unit belongs or the values of the auxiliary variables of other units in the population. The conditional covariance between any two units is zero. This model covers grouping models in which the group formation process is characterised by the auxiliary variables  $z_i$ . The auxiliary variables can be thought of as those variables that determine to which group a unit belongs. More generally, the auxiliary variables can be regarded as the main individual level variables whose distributions are not random across groups because of the choice or migration processes to which the population has been subjected. Contextual variables can also be included in this model as auxiliary variables which take the same value for each unit in the group.

If the vector of auxiliary variables has a marginal distribution with mean  $\mu_z$  and covariance matrix  $\Sigma_{zz}$ , then the marginal mean and covariance matrix of  $y$  are given by  $\mu_y = \mu_{y.z} + \beta'_{yz} \mu_z$  and  $\Sigma_{yy} = \Sigma_{yy.z} + \beta'_{yz} \Sigma_{zz} \beta_{yz}$  respectively. The properties of the sample group level means follow readily from Model B:

$$E[\bar{y}_g | s, z, c] = \mu_y + \beta'_{yz} (\bar{z}_g - \mu_z) \quad (2.22)$$

$$V(\bar{y}_g | s, z, c) = \frac{1}{n_g} \Sigma_{yy.z} \quad (2.23)$$

$$\text{Cov}(\bar{y}_g, \bar{y}_h | s, z, c) = 0 \quad g \neq h. \quad (2.24)$$

The group level statistics then have the following properties

$$E[\bar{y} | s, z, c] = \mu_y + \beta'_{yz} (\bar{z} - \mu_z) \quad (2.25)$$

$$E[S_{yy} | s, z, c] = \Sigma_{yy} + \beta'_{yz} (S_{zz} - \Sigma_{zz}) \beta_{yz} \quad (2.26)$$

$$E[\bar{S}_{yy} | s, z, c] = \Sigma_{yy} + \beta'_{yz} (\bar{S}_{zz} - \Sigma_{zz}) \beta_{yz} \quad (2.27)$$

where  $S_{zz}$  and  $\bar{S}_{zz}$  are defined analogously to  $S_{yy}$  and  $\bar{S}_{yy}$  as given in equation (2.3) and the sentence that follows it.

## 2.4 A Combined Model

The two models considered so far can be thought of as competing explanations of the group effects, but they can be combined into a more realistic model which contains both grouping effects and residual variance components:

**Model C:**

$$E[y_i | z, c] = \mu_{y.z} + \beta'_{yz} z_i \quad (2.28)$$

$$V(y_i | z, c) = \Sigma_{yy.z} \quad (2.29)$$

$$\begin{aligned} \text{Cov}(y_i, y_j | z, c) &= \Delta_{yy.z} \quad \text{if } c_i = c_j \quad i \neq j \\ &= 0 \quad \text{otherwise.} \end{aligned} \quad (2.30)$$

This model allows for group formation processes which are characterised by the auxiliary variables  $z_i$ . It also includes residual within group correlations which reflect random effects which are interpreted as due to unobserved random group level variables after allowing for the grouping variables.

The properties of the sample group level means follow, if the sampling is ignorable given  $(z, c)$  from Model C,

$$E[\bar{y}_g | s, z, c] = \mu_y + \beta'_{yz} (\bar{z}_g - \mu_z) \quad (2.31)$$

and

$$V(\bar{y}_g | s, z, c) = \frac{1}{n_g} (\Sigma_{yy.z} + (n_g - 1) \Delta_{yy.z}) \quad (2.32)$$

$$\text{Cov}(\bar{y}_g, \bar{y}_h | s, z, c) = 0 \quad g \neq h \quad (2.33)$$

$$E[\bar{y} | s, z, c] = \mu_y + \beta'_{yz} (\bar{z} - \mu_z) \quad (2.34)$$

$$\begin{aligned} E[S_{yy} | s, z, c] &= \Sigma_{yy} + \beta'_{yz} (S_{zz} - \Sigma_{zz}) \beta_{yz} \\ &\quad - \frac{\bar{n}^0 - 1}{n - 1} \Delta_{yy.z} \end{aligned} \quad (2.35)$$

$$\begin{aligned} E[\bar{S}_{yy} | s, z, c] &= \Sigma_{yy} + \beta'_{yz} (\bar{S}_{zz} - \Sigma_{zz}) \beta_{yz} \\ &\quad + (\bar{n}^* - 1) \Delta_{yy.z}. \end{aligned} \quad (2.36)$$

Equations (2.17) and (2.18) showed how the effect of aggregation in the variance components model, A, amplifies the contribution of the random group level effects. In equation (2.17) the coefficient of  $\Delta_{yy}$  is  $0(m^{-1})$  whereas in (2.18) it is  $0(\bar{n})$ . For the combined model, C, equations (2.35) and (2.36) show how inclusion of the grouping variables permit the partition of the bias into two additive terms: the first related to the grouping variables, their relationship to the variables of interest and their aggregation effect and the second term involving  $\Delta_{yy.z}$ , the residual components of variance after controlling for the grouping variables. Note that the coefficients of  $\Delta_{yy.z}$  in equations (2.35) and (2.36) are still  $0(m^{-1})$  and  $0(\bar{n})$  respectively as they were in equations (2.17) and (2.18) but the residual components of variance should in general be smaller. The basic assumption in (2.29) is that the residual variance is constant across  $c$ .

The assumption that the sampling is ignorable given  $(z, c)$  means that the sample design can depend on the auxiliary variables and the group indicatives. This allows, for example, the use of stratification based on the values of  $z$  and cluster or multi-stage sampling based on the groups.

The weighted group level matrix  $\bar{S}_{yy}$  is intended to estimate  $\Sigma_{yy}$ . The first bias term in (2.36) is due to the effect of the grouping variables and will be zero if  $\beta_{yz} = 0$  or approximately so if  $\bar{S}_{zz} \doteq \Sigma_{zz}$ . The condition  $\beta_{yz} = 0$  is a strong condition and implies that the variables of interest are unrelated to the grouping variables. The effect of aggregation on the sample covariance of any two variables will depend on the relationships of the variables



with the grouping variables  $z_i$  and we would expect the aggregation effects to be greater for variables more closely related to the grouping variables. The condition  $\bar{S}_{zz} \doteq \Sigma_{zz}$  implies that there are no selection or aggregation effects for the  $z$  variables. These conditions are unlikely to apply in practice and hence bias will result for many variables. The bias due to the sampling and grouping involving the auxiliary variables is determined by  $S_{zz} - \Sigma_{zz}$  for the unit level estimator and by  $\bar{S}_{zz} - \Sigma_{zz}$  for the group level estimator. The term  $\bar{S}_{zz} - \Sigma_{zz}$  reflects the net effect of the sampling and aggregation on the auxiliary variables.

The second bias term in (2.36) will be zero if  $\Delta_{yy.z} = \mathbf{0}$  which implies that, conditional on the grouping variables, there is no residual intra-group correlation among the  $y$  variables. This is unlikely to occur in practice but it is desirable to identify grouping variables that account for as much of the aggregation effects as possible by making this residual term as small as possible.

The effects due to the grouping and sampling depending on  $z$  and the effect due to the residual within group correlation are additive; this will be the case for more complex forms of within group correlations provided the linearity of the model holds. If  $z$  follows a simple variance component model, like Model A then

$$E[\bar{S}_{zz} | s, c] = \Sigma_{zz} + (\bar{n}^* - 1)\Delta_{zz}$$

$$E[\bar{S}_{yy} | s, c] = \Sigma_{yy} + (\bar{n}^* - 1)\beta'_{yz} \Delta_{zz} \beta_{yz} + \Delta_{yy.z} \quad (2.37)$$

and the intra-group covariances of the variables of interest are composed of a component due to the intra-group covariances of the auxiliary variables and the residual components. The right hand side of (2.37) represents a partition of (2.18) since if  $z$  follows a variance components model then so does  $y$  unconditionally. The motivation behind the basic model is to find auxiliary variables so that the residual or conditional within group covariances  $\Delta_{yy.z}$  are small or, ideally, disappear.

## 2.5 Adjusting for Aggregation Effects

Few useful proposals have been made on how to adjust the area level analyses to produce reasonable estimates of the unit level relationship. Duncan and Davis (1953) considered the possible range of the correlation coefficient calculated from a 2 by 2 table with known margins. The resulting bounds are often too wide to be of practical use. Goodman (1959) identified specific conditions for a regression model under which ecological analysis could validly be used to draw inferences regarding relationships at the individual level. Langbein and Litchman (1978) consider some methods that can be applied when grouping is by the

dependent variable and unit level variances are available for both the dependent and all the independent variables in the regression model. However, none of these approaches provide a general approach to the problem.

Examining the bias for  $\bar{S}_{yy}$ , given in (2.36) shows that if we add  $\beta'_{yz}(\Sigma_{zz} - \bar{S}_{zz})\beta_{yz}$  to  $\bar{S}_{yy}$ , the bias term due to the grouping variables would be removed. Now (2.31) implies that

$$E[\bar{B}_{yz} | s, z, c] = \beta_{yz} \quad (2.38)$$

where  $\bar{B}_{yz} = \bar{S}_{zz}^{-1} \bar{S}_{zy}$ .

If the covariance matrix of  $z$ ,  $S_{zzs_0}$ , from a unit level sample  $s_0$  drawn from  $m_0$  groups was available then the adjusted estimator

$$\hat{\Sigma}_{yy}(z) = \bar{S}_{yy} + \bar{B}'_{yz}(S_{zzs_0} - \bar{S}_{zz})\bar{B}_{yz} \quad (2.39)$$

should remove the aggregation bias due to the grouping variables  $z$ , provided  $S_{zzs_0}$  is close to  $\Sigma_{zz}$ . The source for  $S_{zzs_0}$  may be quite independent of the data used in  $\bar{S}_{yy}$  and  $\bar{B}_{yz}$ . Steel (1985) shows that the adjusted estimator (2.39) can be obtained as the MLE of  $\Sigma_{yy}$  (with the usual replacement of  $m - 1$  by  $m$  etc.). If normality of the distribution of  $(y, z)$  applies,  $s_0$  is a simple random sample from the population and  $\Delta_{yy.z} = \mathbf{0}$ . The adjusted estimator corresponds to the Pearson (1903) adjustment considered by Holt, Smith and Winter (1980) in the case of regression analysis and Smith and Holmes (1989) in the case of multivariate analysis. In these cases the adjustment is applied to statistics calculated from unit level data obtained from a sample whose design depends on the auxiliary variables. In our case the adjustment is applied to statistics calculated from area means and the auxiliary variables used in the adjustment include grouping variables as well as any design variables. The adjusted estimator of  $\mu_y$  is

$$\hat{\mu}_y(z) = \bar{y} + \bar{B}'_{yz}(\bar{z}_{s_0} - \bar{z}) \quad (2.40)$$

where  $\bar{z}_{s_0}$  is the mean calculated from  $s_0$ .

From (2.34) and (2.38) we see that

$$E[\hat{\mu}_y(z) | s, z, s_0, c] = \mu_y + \beta'_{yz}(\bar{z}_{s_0} - \mu_z). \quad (2.41)$$

Moreover, Steel (1985) shows that (2.36) and (2.38) imply

$$E[\hat{\Sigma}_{yy}(z) | s, z, s_0, c] = \Sigma_{yy} + \beta'_{yz}(S_{zzs_0} - \Sigma_{zz})\beta_{yz} + (\bar{n}^* - 1)\Delta_{yy.z} + O(m^{-1}) \quad (2.42)$$

provided  $\text{tr}(\bar{S}_{zz}^{-1} S_{zzs_0})$  and  $\bar{n} \text{tr}((\bar{S}_{zz}^{-1} S_{zzs_0} - I) \bar{S}_{zz}^{-1} \bar{S}_{zz}^{(2)})$  are bounded, where  $\bar{S}_{zz}^{(2)}$  is defined similarly to  $\bar{S}_{zz}$  with  $n_g$  replaced by  $n_g^2/\bar{n}$ .

Comparing (2.42) with (2.35) we see that the component of bias due to the grouping variables has been adjusted to that associated with the use of  $S_{yys_0}$ , if it had been available. The estimator adjusts for the aggregation effects that have acted through  $z$ . It also adjusts the effect of the sampling design from that associated with  $s$  to that associated with  $s_0$ .

Suppose that the sampling design used to generate  $s_0$  and the values of the auxiliary variables are generated from a superpopulation such that

$$E[\bar{z}_{s_0} | s_0, c] = \mu_z + 0(m_0^{-1}) \quad (2.43)$$

$$E[S_{zzs_0} | s_0, c] = \Sigma_{zz} + 0(m_0^{-1}) \quad (2.44)$$

where  $m_0$  is the number of groups in  $s_0$ .

In such cases

$$E[\hat{\mu}_y(z) | s, s_0, c] = \mu_y + 0(m_0^{-1}) \quad (2.45)$$

$$E[\hat{\Sigma}_{yy}(z) | s, s_0, c] = \Sigma_{yy} + (\bar{n}^* - 1)\Delta_{yy,z} + 0(\tilde{m}^{-1}) \quad (2.46)$$

where

$$\tilde{m} = \min(m, m_0).$$

Conditions (2.43) and (2.44) would apply if the population  $z$  values across groups arose from a variance component model similar to model A and the sampling design for  $s_0$  depended only on the grouping but not any auxiliary variables. Sampling designs such as simple random sampling or equal probability cluster or multi stage sampling fulfil this condition. Use of census data, so that  $s_0$  is the entire finite population is also applicable.

It is thus possible to adjust for the bias due to the grouping variables provided some unit level sample covariance matrix for  $z$  is available. The motivation for the approach is a situation where the predominant group effects can be attributed to selectivity or grouping effects acting through the grouping variables. The adjustment for the auxiliary variables removes the effect of the apparent intra-group correlation due to these variables. The adjusted estimator still has a component of bias due to  $\Delta_{yy,z}$  and if  $z$  is not effective in significantly reducing the intra-group correlations then this term can still be important. This approach therefore relies on choice of appropriate auxiliary variables to reduce the intra-group correlations.

If the sampling design for  $s_0$  and the superpopulation model for  $z$  are such that (2.43) and (2.44) do not apply then  $\bar{z}_{s_0}$  and  $S_{zzs_0}$  can be replaced by estimators  $\hat{\mu}_{zs_0}$  and  $\hat{\Sigma}_{zzs_0}$  in the calculation of the adjusted estimators  $\hat{\mu}_y(z)$

and  $\hat{\Sigma}_{yy}(z)$ . The resulting expectations of the adjusted estimators are given by (2.41) and (2.42) with  $\bar{z}_{s_0}$  replaced by  $\hat{\mu}_{zs_0}$  and  $S_{zzs_0}$  replaced by  $\hat{\Sigma}_{zzs_0}$ . There are a number of choices available for the estimators  $\hat{\mu}_{zs_0}$  and  $\hat{\Sigma}_{zzs_0}$  calculated from the sample  $s_0$ . Smith and Holmes (1989) consider a range of model based and design based estimators that can be used. For example suppose the sample design used to obtain  $s_0$  involved stratification according to the values of the vector of size variables  $x$ . Denote the sample inclusion probability for population unit  $i$  as  $\Pi_i$  and the associated probability based weight is  $w_i = (\Pi_i)^{-1}$ . The probability weighted estimator of  $\mu_z$  is  $\bar{z}_{s_0}^* = \sum_{i \in s_0} w_i z_i$ , and of  $\Sigma_{zz}$  is  $S_{zzs_0} = \sum_{i \in s_0} w_i z_i z_i' - w_0^{-1} \bar{z}_{s_0}^* \bar{z}_{s_0}^{*'} where  $w_0 = \sum_{i \in s_0} w_i$ .$

The Pearson based adjusted estimators of  $\mu_z$  and  $\Sigma_{zz}$  are  $\bar{z}_{s_0} + B'_{zxs_0}(\bar{x}_u - \bar{x}_{s_0})$  and  $S_{zzs_0} + B'_{zxs_0}(S_{xxu} - S_{xxs_0}) B_{zxs_0}$  respectively. Here  $\bar{x}_u$  and  $S_{xxu}$  are the mean vector and covariance matrix of the design variables in  $x$  in the finite population and  $B_{zxs_0} = S_{xxs_0}^{-1} S_{xzs_0}$ .

Probability weighted Pearson based adjusted estimates may also be considered, i.e.,  $\bar{z}_{s_0}^* + B'_{zxs_0}(\bar{x}_u - \bar{x}_{s_0}^*)$  and  $S_{zzs_0}^* + B'_{zxs_0}(S_{xxu} - S_{xxs_0}^*) B_{zxs_0}^*$ .

Here  $\bar{x}_{s_0}^*$  and  $S_{xxs_0}^*$  are defined analogously to  $\bar{z}_{s_0}^*$  and  $S_{zzs_0}^*$  respectively and  $B_{zxs_0}^* = S_{xxs_0}^{*-1} S_{xzs_0}^*$ . The approach taken so far is strongly model based and so model based estimators of  $\mu_y$  and  $\Sigma_{zz}$  would be preferred. However, in practice the data available for use in the adjustment may comprise published  $p$ -weighted estimators of means and covariances obtained from the sample  $s_0$ , which is independent of  $s$ . Thus

$$E_{p_0}[\hat{\mu}_{zs_0} | z, c] = \bar{z}_u$$

$$E_{p_0}[\hat{\Sigma}_{zzs_0} | z, c] = S_{zzu}$$

where  $\bar{z}_u$  and  $S_{zzu}$  are the mean vector and covariance matrix of the auxiliary variables in the finite population and  $E_{p_0}$  represents the expectation with respect to repeated sampling using the sampling design employed to obtain  $s_0$ , i.e., the randomization distribution. Thus from (2.41) and (2.42)

$$E[\hat{\mu}_y(z) | s, z, c] = \mu_y + \beta_{yz}'(\bar{z}_u - \mu_z)$$

$$E[\hat{\Sigma}_{yy}(z) | s, z, c] = \Sigma_{yy} + \beta_{yz}'(S_{zzu} - \Sigma_{zz})\beta_{yz} + (\bar{n}^* - 1)\Delta_{yy,z} + 0(m^{-1}).$$

These expectations are taken over the statistical model generating the  $y$  values and the randomization distribution associated with  $s_0$ . In practice  $\bar{z}_u$  and  $S_{zzu}$  will be very close to  $\mu_z$  and  $\Sigma_{zz}$  respectively.



### 3. IDENTIFYING GROUPING VARIABLES

In the previous section we introduced a set of auxiliary variables,  $z$ , which characterised the area differences and which were used to adjust the aggregated analysis to reduce the aggregation bias. If the auxiliary variables were totally successful then  $\Delta_{yy,z}$  would be reduced to zero and the adjustment method would remove the aggregation bias completely. In practice the auxiliary variables for which  $\Delta_{yy,z} = 0$  are unknown. Also we will be restricted to sets of variables for which area level means are available as part of the data set under analysis and for which an estimate  $\hat{\Sigma}_{zz}$  of the unit level covariance matrix is available. Basic demographic information and housing variables commonly available from the Census may be used. However these variables may not fully characterise the grouping process and so they may not explain as much of the between area difference as we might wish.

#### 3.1 An Analysis Strategy

In practice the grouping variables will not be known. We need a strategy for identifying adjustment variables for which an estimate of the unit level covariance matrix is available and which account for group effects. One strategy involves the following steps:

- 1) Identify a set of variables that cover the same subject area as the variables of interest, but for which both area level and unit level data are available for some period in the past. Previous Census data may be suitable.
- 2) Add to this set, variables (such as demographic and housing variables) which are candidate  $z$  variables since they are known to be strongly associated with area differences. Estimates of both the area level and unit level covariance matrices must also be available for the same period in the past.
- 3) Carry out an analysis of these data to identify the variables which account most strongly for the area level effects among the variables of interest. This analysis, which we term a CGV analysis, will be described below.
- 4) Identify from (3) a set of adjustment variables which are available within the current data set and for which the current unit level covariance matrix is available from some source.
- 5) For some variables of interest it may be possible to obtain estimates of unit level variances or covariances, from published tables for example. From these calculate aggregation effects  $\bar{Q}_{aa} = \bar{s}_{aa}/s_{aa}$  or  $\bar{Q}_{ab} = \bar{s}_{ab}/s_{ab}$ .
- 6) Use the variables identified in (4) to adjust the aggregate analysis for the variables of interest and check the adjusted aggregation effects corresponding to (5) to monitor the success of the adjustment.

#### 3.2 The Ideal Grouping Variables

We first consider the ideal set of grouping variables that could be used for adjustment so as to identify the appropriate (CGV) analysis that could be followed for the analysis of aggregated data using the strategy outlined above.

Let us suppose that for the complete set of variables of interest we have the area level variance-covariance matrix  $\bar{S}_{yy}$  and the unit level variance-covariance matrix  $S_{yys_1}$  based on a sample  $s_1$ . Of course if this occurred in practice the aggregation problem would disappear since we could discard  $\bar{S}_{yy}$  and simply use  $S_{yys_1}$ , as an estimate of  $\Sigma_{yy}$ . However there are three reasons for considering this situation. Firstly it helps to throw light on the grouping structure which determines the relationship between  $\bar{S}_{yy}$  and  $S_{yys_1}$ . Secondly it may be that  $\bar{S}_{yy}$  and  $S_{yys_1}$  are available at some point in time such as census day but that further analysis of a new version of  $\bar{S}_{yy}$  is to be based on inter-censal data when  $S_{yys_1}$  is unavailable. If the grouping structure persists over time, as we might expect, then the analysis of the census day versions of  $\bar{S}_{yy}$  and  $S_{yys_1}$  might help the subsequent inter-censal analysis by identifying the key variables that explain a large proportion of the aggregation effects. These possibilities underpin the strategy outlined in section 3.1 above. Thirdly if the variables in  $y$  cover a large range of socio-economic and demographic variables, as occurs in the census, then the key variables that account for the grouping effects for the variables may also explain much of the grouping effects of other socio-economic and demographic variables. Note that the two samples  $s$  and  $s_1$  may be identical but in general do not need to be. For example  $s$  may correspond to an administrative source which is effectively a census that provides aggregate data for geographic areas, and  $s_1$  is a sample survey from which individual level data are made available without any geographic identifiers.

To help identify the important variables associated with the grouping Steel (1985) suggests that  $\hat{\theta}_1, \dots, \hat{\theta}_p$ , the eigenvalues of  $S_{yys_1}^{-1} \bar{S}_{yy}$ , be calculated as well as the matrix  $\hat{D}_y = [\hat{d}_1, \dots, \hat{d}_p]$  such that

$$\hat{D}_y' \bar{S}_{yy} \hat{D}_y = \text{diag}(\hat{\theta}_k) \quad \text{and} \quad \hat{D}_y' S_{yys_1} \hat{D}_y = I.$$

The variables defined by the transformation

$$\hat{u}_i = \hat{D}_y' y_i$$

successively have maximum ratio of between group to sample total variance and have zero sample correlation at the unit and group level and unit level sample variance of 1. These variables are called the sample Canonical Grouping Variables (CGVs). The sample CGVs have the maximum intra-group correlation. Note that  $\text{tr}(S_{yys_1}^{-1} \bar{S}_{yy}) = \sum_k \hat{\theta}_k$  can be defined as the multivariate aggregation effect.

Note that the matrix  $\hat{D}_y$  will exist even if  $S_{yys_1}$  and  $\bar{S}_{yy}$  are based on different samples so long as the former is positive definite and the latter is positive semi-definite. Furthermore the variances of the CGV's will be non-negative. However, when  $s$  and  $s_1$  are distinct it is possible that the maximum variance of a CGV could exceed  $(N-1)/(M-1)$  which is the maximum possible aggregation effect. In this case the CGV has an implied negative within group variance component. For our purposes this may not matter since we are interested in identifying important grouping variables but in principle the offending variance of the CGV could be set to its theoretical maximum. The sample CGVs are obtained from the eigenvectors of  $A_{yy} = S_{yys_1}^{-1} \bar{S}_{yy}$ . If  $s$  and  $s_1$  are the same sample then  $A_{yy}$  is the sample regression coefficient for the regression of the group level means on the unit level values calculated over the unit level sample. In this case the sample CGVs are in fact the sample canonical variates relating the unit level and group level data and  $\hat{\theta}_k$  are the sample canonical correlations.

Having calculated the CGVs the difference between the sample group level and unit level covariance matrix can be expressed as

$$\bar{S}_{yy} - S_{yys_1} = \sum_k (\hat{\theta}_k - 1) \hat{\phi}_k \hat{\phi}_k'$$

where  $\hat{\phi}_k$  is the vector of sample covariances between the  $k$ -th CGV and the original variables. Hence the difference between the group level and unit level covariance matrix can be partitioned into  $k$  orthogonal elements, one for each CGV.

For the covariance between  $y_{ia}$  and  $y_{ib}$ , the difference between the sample group level covariance,  $\bar{s}_{ab}$  and unit level covariance  $s_{ab}$  (where  $\bar{s}_{ab}$  and  $s_{ab}$  elements of  $\bar{S}_{yy}$  and  $S_{yys_1}$ , respectively) is

$$\bar{s}_{ab} = s_{ab} + (s_{aa}s_{bb})^{1/2} \sum_k (\hat{\theta}_k - 1) \hat{\rho}_{ak} \hat{\rho}_{bk}$$

where  $\hat{\rho}_{ak} = \hat{\phi}_{ak}/s_{aa}^{1/2}$  is the sample correlation between the  $a$ -th variable and the  $k$ -th sample CGV.

If the first  $q$  sample CGVs are used to calculate an adjusted group level variance matrix, *i.e.*,  $\hat{u}_{qi} = \hat{D}_q' y_i$  where  $\hat{D}_q = [\hat{d}_1, \dots, \hat{d}_q]$ , are used as the auxiliary variables

$$\hat{\Sigma}_{yy}(\hat{u}_q) = \bar{S}_{yy} + \bar{B}_{yu_q}' (S_{u_q u_q s_0} - \bar{S}_{u_q u_q}) \bar{B}_{yu_q}$$

then the first  $q$  terms of the decomposition are removed *i.e.*,

$$\hat{\Sigma}_{yy}(\hat{u}_q) = S_{yys_1} + \sum_{k=q+1}^p (\hat{\theta}_k - 1) \hat{\phi}_k \hat{\phi}_k'$$

and  $\text{tr}(S_{yys_1}^{-1} \hat{\Sigma}_{yy}(\hat{u}_q)) = \sum_{k=q+1}^p \hat{\theta}_k$ . In fact use of the first  $q$  CGVs provides the matrix of rank  $q$  that minimizes  $\|S_{yys_1} - \hat{\Sigma}_{yy}(\hat{u}_q)\|$ . Hence by examining the quantities

$$\sum_{k=q+1}^p \hat{\theta}_k \quad \text{and} \quad 1 + \sum_{k=q+1}^p (\hat{\theta}_k - 1) \hat{\rho}_{ak}^2$$

$$\text{for } q = 0, \dots, p-1$$

it is possible to examine how the proportion of the overall aggregation effect and the aggregation effect for each variable can be explained by the first  $q$  sample CGVs.

The preceding analysis will suggest how many dimensions are required to effectively explain and hence remove a specified amount of the aggregation effects. Moreover by looking at the loadings of the original variables in the CGVs, it should be possible to identify which variables play the major role in ‘‘explaining’’ the aggregation effects of the other variables. It is these variables that researchers should concentrate on obtaining unit level data for, to use in the adjusted estimator.

These results have some important implications for the use of group level data supplemented by limited unit level data, since they open the way to combining sample survey data and group level data from one or more sources and suggest a strategy for the analysis of group effects and group level data.

#### 4. SOME EMPIRICAL RESULTS

We illustrate the ideas of the previous sections with an analysis of the 1991 UK population census data for the Local Authority District (LAD) of Reigate, Banstead and Tandridge. The LAD population is 188,700 people contained in 371 EDs giving an average number of people per ED of  $\bar{n} = 508.6$ . Group level data are available on a complete count basis for each ED in the LAD from the Small Area Statistics (SAS) data file. Corresponding unit level data for the LAD are obtained from a 2 per cent Sample of Anonymized Records of individuals (SAR). The records in the SAR cannot be identified with any specific ED within the LAD thus in this situation we have  $\bar{S}_{yy}$  based upon complete data for each ED from the SAS and we have an estimate of  $\bar{S}_{yys_1}$  based on a 2 percent sample from the SAR. The following analysis is based upon 16 census variables for each person.

For each variable the group level data and the unit level data were used to calculate the aggregation effect,  $\bar{Q}_a = \bar{s}_{aa}/s_{aa}$ . The parameter  $\delta_{aa} = \Delta_{aa}/\Sigma_{aa}$ , defined on the appropriate diagonal elements of  $\Delta_{yy}$  and  $\Sigma_{yy}$  is the intra-group correlation for the  $a$ -th variable. An estimate  $\hat{\delta}_{aa}$  of the intra-group correlation can be obtained from (2.18) since  $\bar{Q}_a = 1 + (\bar{n}^* - 1) \hat{\delta}_{aa}$ . The results for the variables are given in Table 1. The intra-group



**Table 1**  
Aggregation Effects and Intra-class Correlations for  
Census Variables in Reigate LAD

	Aggregation Effect	Intra-class Correlation
Persons aged 18-29	9.20	.016
Persons aged 30-44	4.56	.007
Persons aged 45-59*	5.97	.010
Persons aged 60 and over*	17.17	.032
Female	1.08	.000
Non-white*	8.29	.014
Married	6.24	.010
Limiting long term illness	7.24	.012
Persons employed full time	8.55	.015
Persons unemployed	2.27	.003
Other employment status	11.19	.020
Head of h'hold born UK	4.48	.007
Head of h'hold born New Commonwealth	3.59	.005
Migrant head of household	9.04	.016
≤ 1.5 persons per room: density	27.96	.053
Persons in 0 car households	32.98	.063

\* Selected for adjustment variables.  
Source: Reigate and Banstead; Tandridge LAD 1991 census data.

correlations are generally small but the number of observations in each ED implies that the aggregation effects can be high (see the comment following equation (2.18)).

Figure 1a shows a plot of the group level correlation,  $\bar{r}_{ab}$ , against the individual level correlation,  $r_{ab}$ , for every pair of variables. Note the strong aggregation effects which are revealed through the characteristic S-shaped plot. Small correlations at the unit level are generally magnified so that for most cases  $|\bar{r}_{ab}|$  is much larger than  $|r_{ab}|$ .

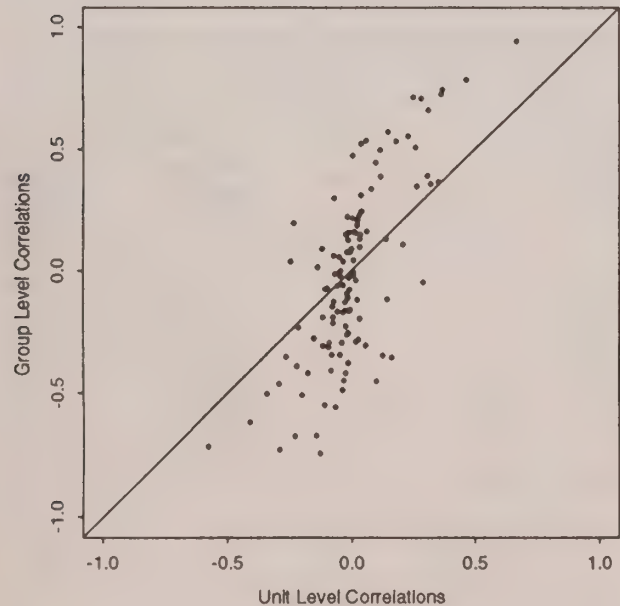


Figure 1a.

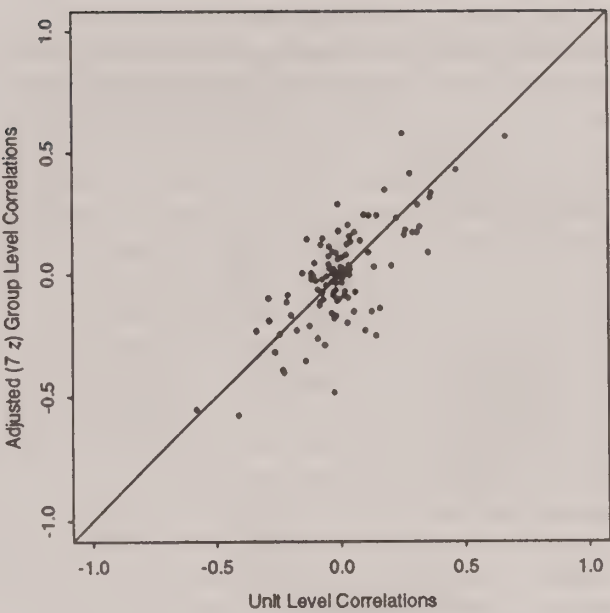


Figure 1b.

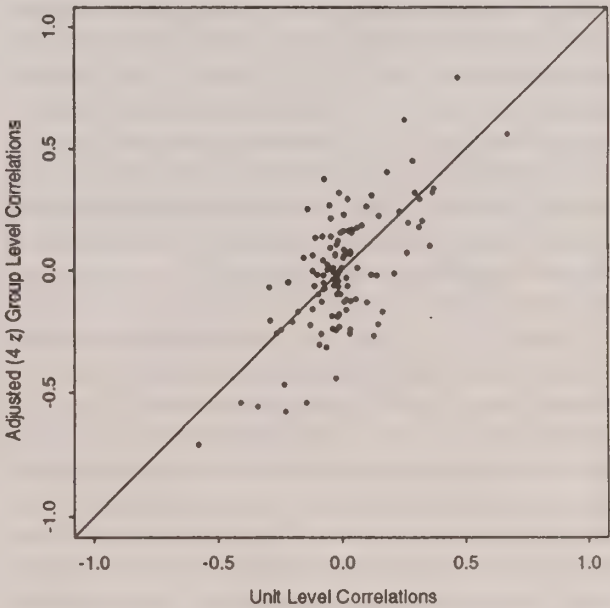


Figure 1c.

Since in this case we have  $\bar{S}_{yy}$  and  $S_{yys_1}$  we may carry out a canonical grouping variable analysis so as to understand the more important features of the grouping structure. Table 2 shows the loadings on the 16 variables for the first five canonical grouping variables which together account for 89% of the multivariate aggregation effect.

The first CGV has high loadings on high density occupation and car (*i.e.*, auto) access and might be interpreted as a socio-economic factor. The second CGV has high loadings the variables indicating people in the two oldest age groups. It is noticeable, also, that the proportion of

**Table 2**  
First Five CGV's for Variables in Table 1

	CGV1	CGV2	CGV3	CGV4	CGV5
Persons aged 18-29	0.4	0.3	0.9	1.1	0.1
Persons aged 30-44	0.1	0.5	0.36	1.0	0.2
Persons aged 45-59*	-0.1	1.2	-0.2	1.0	0.1
Persons aged 60 and over*	0.3	2.2	-0.5	2.6	0.9
Female	0.1	0.0	0.0	0.3	0.1
Non-white*	0.5	-0.4	1.4	-1.1	5.2
Married	-0.2	-0.5	-0.4	-0.8	-0.1
Limiting long term illness	0.3	0.1	-0.2	0.2	0.3
Persons employed full time	0.7	-0.3	0.2	1.2	0.4
Persons unemployed	0.7	0.0	-0.1	0.0	-0.4
Other employment status	0.1	0.1	0.0	-0.2	-0.1
Head of h'hold born UK	0.5	-0.1	-1.0	0.4	0.2
Head of h'hold born New Commonwealth	0.0	-0.1	-0.3	0.1	0.6
Migrant head of household	0.2	0.1	1.4	0.6	-1.3
≤ 0.5 persons per room	-1.4	0.3	1.2	-0.7	-0.2
Persons in 0 car households	2.2	0.6	0.8	-1.9	-0.7

\* Selected for adjustment variables.

Source: Reigate and Banstead; Tandridge LAD 1991 census data.

non-white heads of household contributes to the later CGV's. As might be expected, variables such as proportion Female, that exhibit almost no intra-group correlation and hence no aggregation effect make virtually no contribution to the CGV's. Such variables do not vary across areas and hence generally have no explanatory power.

In usual practice a CGV analysis will not be possible since if  $S_{yy}$  was available there would usually be no need to carry out an aggregate analysis. However the CGV analysis suggests variables that may be important since they load highly on the first few CGVs.

It is well known in the UK context that housing tenure variables (which are not contained in the 16 variables of interest) have a powerful association with a wide variety of socio-economic, attitudinal and health variables. There are strong reasons for assuming that using these as auxiliary,  $z$ , variables for adjustment would account for a substantial proportion of the first socio-economic dimension and may act in place of the density of occupancy and car access variables that are seen to be important for the first CGV. The other reason for considering those variables is that if the present analysis is to act as an illustration of what might be achieved in other situations then basic tenure and housing variables are more likely to be available as adjustment variables than density of occupation and car access. In the light of the CGV analysis and in the spirit of identifying a small number of adjustment variables which could be expected to be available in many situations, we identify a set of seven potential adjustment variables. These are the three variables of interest identified in Table 1 identified by an asterisk (Age 45-59, Age 60+, non-white) and the four housing variables listed in Table 3 together with their aggregation effects and intra-cluster correlations.

**Table 3**  
Aggregation Effects and Intra-class Correlations for Household Level Variables in Reigate LAD

Variable		Aggregation Effect	Intra-class Correlation
Tenure:	LA Rented	133.43	0.261
	Owner Occupier	90.83	0.177
Stock:	Det/semi/terrace	90.03	0.175
	Good Amenities	59.52	0.113

Source: Reigate and Banstead; Tandridge LAD 1991 census data.

In what follows the group level covariance matrix for the original 16 variables will be adjusted by the unit level covariance matrix for 7  $z$ -variables (three of the basic demographic variables in the original set and four household variables).

Two overall measures of the effectiveness of the adjustment were calculated. The first is

$$1 - \frac{\text{tr}(S_{yys_1}^{-1} \hat{\Sigma}_{yy}(z)) - 1}{\text{tr}(S_{yys_1}^{-1} \bar{S}_{yy}) - 1}$$

which is the reduction in the multivariate aggregation effect and the second is

$$\frac{\|S_{yys_1} - \bar{S}_{yy}\| - \|S_{yys_1} - \hat{\Sigma}_{yy}(z)\|}{\|S_{yys_1} - \bar{S}_{yy}\|}$$

which shows the reduction in the generalised distance between the unit level and group level covariance matrices before and after adjustment.

**Table 4**

Z-variable Combination	No. of Variables	% reduction in	
		Multivariate Aggregation Effect	Generalised Distance
60+	1	16	24
45-59, 60+ Tenure Stock	2	38	53
	2	30	21
	2	31	19
45-59, 60+, NW	3	44	54
45-59, 60+, tenure 45-59, 60+, stock	4	57	71
	4	57	69
45-59, 60+, tenure, NW 45-59, 60+, stock, NW	5	63	72
	5	62	70
45-59, 60+, stock, tenure, NW	7	68	75



Table 4 shows the effect of using various combinations of variables for adjustment of the aggregated analysis. The two age variables are clearly important (accounting for 38% of the multivariate aggregation effect and 53% of the generalized distance) but the Tenure or Housing Stock variables are also important. When Tenure or Housing Stock are used in conjunction with age the percentage reduction in either measure is close to the sum of the effects of the variables separately showing that age and Tenure or Housing Stock are acting as distinct adjustment variables. Obviously the greatest success is achieved by including all 7 adjustment variables and accounts for 68% and 75% respectively of the two aggregation measures.

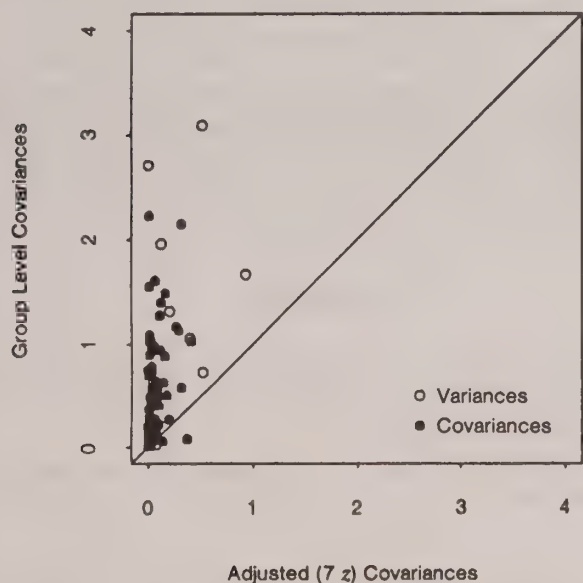


Figure 2a.

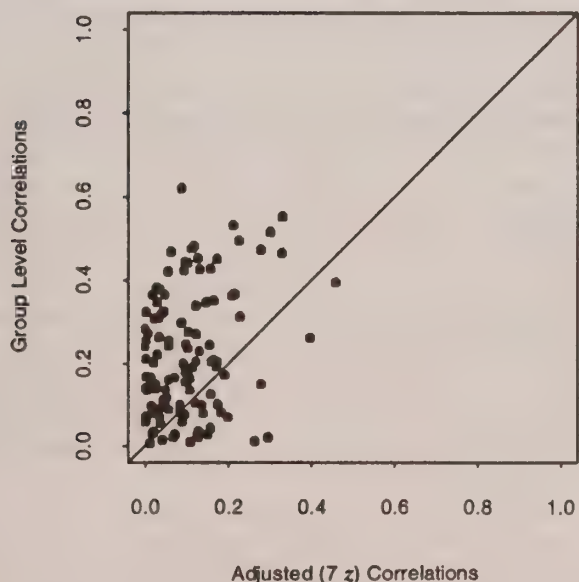


Figure 2b.

These results show that around 70% of the aggregation effects have been removed by the adjustment. Figures 2a and 2b show the effect of adjustment by these variables. In Figure 2a the vertical axis contains  $|\bar{s}_{ab} - s_{abs_1}|$ , the absolute bias for the group level covariance for each pair of variables. The horizontal axis contains  $|\hat{\Sigma}_{ab}(z) - s_{abs_1}|$ , the absolute bias of the adjusted estimator. The hollow symbol is used for variances of the  $y$  variables, and the solid symbol is used for covariances. Almost all of the plotted values show that the biases after adjustment are smaller (often much smaller) than the original bias. In almost all cases the adjustment has had a substantial improvement. Figure 2b shows the corresponding plot for correlations rather than covariances. (Correlations of  $y_a, y_a$  have obviously been omitted from this plot.) Again there is a strong improvement with the residual bias after adjustment being much smaller than the original bias for the group level analysis. The results are not as successful as for the covariances, since in some cases small biases for the group level analysis have been made worse. In this case the adjustments are applied to the covariance and the two variances used in each correlation coefficient. There is more potential for the relative changes in each component to lead to a correlation which is worse than the original. However, almost all of the large biases at the group level have been improved.

Figure 1b shows the plot of the adjusted group level correlations,  $\bar{r}_{ab}(z)$ , obtained from  $\hat{\Sigma}_{yy}(z)$  against the unit level correlations and can be compared with the original unadjusted plot in Figure 1a. The characteristic S-shaped curve shown in Figure 1a has been replaced by a plot of points which lie about the line  $\bar{r}_{ab}(z) = r_{ab}$  as we would want if aggregation bias is removed.

Figures 1b, 2a and 2b show that a substantial reduction to the aggregation effect can be achieved by using 4 housing variables and 3 of the original  $y$  variables. This implies adjusting the original 120 variances and covariances in the  $16 \times 16$  matrix by 21 variances and covariances for the  $z$  variables. As an illustration of what might be achieved with minimal information we reduce the adjustment variables to the four involving age and Tenure. From Table 4 we see that these account for 57% and 71% of the two measures of aggregation. Figures 3a and 3b show the corresponding plots to Figures 2a and 2b for this case. Figure 1c shows the plot of the adjusted correlations using 4 variables against the individual level correlations. Obviously the adjustment is not as successful but it is encouraging to see what can be achieved with so few adjustment variables. As a further measure of the effect of the adjustment the median absolute difference between  $\bar{r}_{ab}$  and  $r_{ab}$  was 0.186. After adjusting by 4 variables this was reduced to 0.126 and after adjusting 7 variables to 0.090. The corresponding median values for  $|\bar{s}_{ab} - s_{ab}|$  were 0.173, 0.039 and 0.017 respectively.

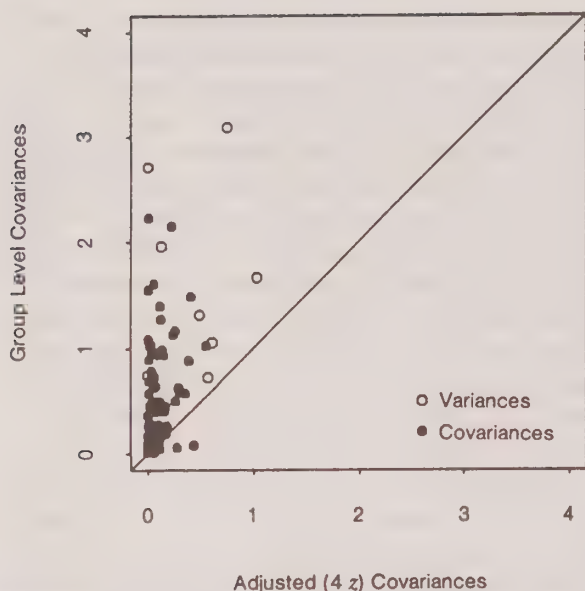


Figure 3a.

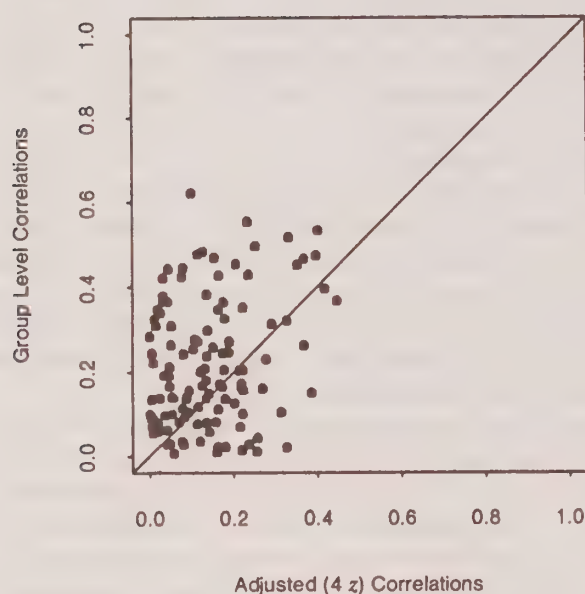


Figure 3b.

## 5. CONCLUSIONS AND DISCUSSION

A model for grouped populations has been proposed which leads to a decomposition of the bias observed in group level analysis based on covariance matrices into two components. The first component is due to the grouping variables and the second is due to the residual intra-group correlations between the  $y$  variables given the grouping variables  $z$ . This decomposition provides an understanding of the magnitude of aggregation effects. It also provides a way of removing the bias due to the grouping variables if additional information about the unit level covariance matrix of the grouping variables is available.

In many countries there are many group level data available at different levels of aggregation from the census and many other sources. The development of Geographic Information Systems will increase the availability of such data. It is important to analyse and decompose the group effects and the theory developed and the strategy proposed here provide a framework for achieving this. A proper understanding of which variables explain most of the group effects, and therefore should be used in adjusting ecological analyses, will open the way to making use of aggregated data.

## ACKNOWLEDGEMENTS

This research was supported by Grant Number H507 26 5013 from the Economic and Social Research Council, United Kingdom. The authors also acknowledge gratefully helpful comments from the associate editor and referees.

## REFERENCES

- ARBIA, G. (1989). *Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems*. Dordrecht: Kluwer.
- BLALOCK, H.M. (1964). *Causal Inference in Nonexperimental Research*. Chapel Hill NC: University of North Carolina Press.
- BLALOCK, H.M. (1979). Measurement and conceptualization problems: The major obstacle to integrating theory and research. *American Sociological Review*, 44, 881-894.
- BLALOCK, H.M. (1985). Cross level analysis. In *The Collection and Analysis of Community Data*, (Ed. J.B. Casterlin), ISI, World Fertility Survey.
- CLARK, W.A.V., and AVERY, K.L. (1976). The effect of data aggregation in statistical analysis. *Geographical Analysis*, 8, 428-438.
- DUNCAN, D.P., and DAVIS, B. (1953). An alternative to ecological correlation. *American Sociological Review*, 18, 665-666.
- FOTHERINGHAM, A.S., and WONG, D.W.S. (1991). The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning, A*, 23, 1025-1044.
- GEHLKE, C.E., and BIEHL, K. (1934). Certain effects of grouping upon the size of the correlation coefficient in census tract material. *Journal of the American Statistical Association*, 29, Supplement, 169-170.
- GOODMAN, L.A. (1959). Some alternatives to ecological correlation. *American Journal of Sociology*, 64, 610-625.
- HANNAN, M.T., and BUSTEIN, L. (1974). Estimation from grouped observations. *American Sociological Review*, 39, 374-392.



- HOLT, D., SMITH, T.M.F., and WINTER, P.D. (1980). Regression analysis of data from complex surveys. *Journal of the Royal Statistical Society, A*, 143, 474-87.
- HOLT, D., and SCOTT, A.J. (1981). Regression analysis using survey data. *The Statistician*, 30, 169-173.
- LICHTMAN, A.J. (1974). Correlation, regression, and the ecological fallacy: A critique. *Journal of Interdisciplinary History*, 4, 417-433.
- LANGBEIN, L.I., and LICHTMAN, A.J. (1978). *Ecological Inference*. Thousand Oaks, CA: Sage.
- OPENSHAW, S. (1984). Ecological fallacies and the analysis of areal census data. *Environment and Planning, A*, 6, 17-31.
- OPENSHAW, S., and TAYLOR, P.J. (1979). A million or so correlation coefficients: three experiments on the modifiable areal unit problem. In *Statistical Applications in the Spatial Sciences*, (Ed. N. Wrigley), 127-144.
- PEARSON, K. (1903). On the influence of natural selection on the variability and correlation of organs. *Philosophical Transactions of the Royal Society, A*, 200, 1-66.
- PERLE, E.D. (1977). Scale changes and impacts on factorial ecology structures. *Environment and Planning, A*, 9, 549-558.
- RAO, C.R. (1973). *Linear Statistical Inference and its Applications*, (2nd Ed.). New York: Wiley.
- ROBINSON, W.S. (1950). Ecological correlations and the behaviour of individuals. *American Sociological Review*, 15, 351-357.
- SMITH, K.W. (1977). Another look at the clustering perspective on aggregation problems. *Sociological Methods and Research*, 5, 289-316.
- SMITH, T.M.F., and HOLMES, D. (1989). Multivariate analysis. In *Analysis of Complex Surveys*, (Eds. C.J. Skinner, D. Holt and T.M.F. Smith), 165-187.
- STEEL, D. (1985). Statistical Analysis of Populations with Group Structure. Unpublished PhD Thesis, Department of Social Statistics, University of Southampton.
- STEEL, D., and HOLT, D. (1994). Analysing and Adjusting Aggregation Effects: The Ecological Fallacy Revisited. Department of Applied Statistics, University of Wollongong, Preprint 1/94.
- STEEL, D., and HOLT, D. (1995). Rules for random aggregation. *Environment and Planning* (to appear).
- YULE, U., and KENDALL, M.S. (1950). *An Introduction to the Theory of Statistics*. Glendale, CA: Griffin.





# Linearization Methods for Single Phase and Two-Phase Samples: A Cookbook Approach

DAVID A. BINDER<sup>1</sup>

## ABSTRACT

There are a number of asymptotically equivalent procedures for deriving the Taylor series approximation of variances for complex statistics. In Binder and Patak (1994) the theoretical justification for one class of methods was derived. However, many of these methods can be derived for practical examples using straightforward techniques that are not clearly described in Binder and Patak. In this paper we give a "cookbook" approach that can be used for many examples, and that has been shown to have good finite sample properties. Normally the method of choice becomes clear through arguments such as model-assisted methods or linearizing the jackknife; however, using our approach yields the desired results more directly. As well, we present new results on the application of these techniques to two-phase samples.

**KEY WORDS:** Complex surveys; Variance estimation; Ratio estimator; Regression estimator; Wilcoxon rank sum test; Estimating equations.

## 1. THE METHOD

The derivation of the asymptotic variance for a wide class of estimators from complex survey samples is now well established in the literature, at least to a first order approximation. However, there are a number of competing estimators of the variance, all of which are asymptotically equivalent. In this paper, we discuss a simple derivation of one of the most favoured of these estimators in a general setting. This simple derivation is useful for practitioners, who may be baffled by the choices available, and need a quick solution to the problem.

We start with a simple example of the approach using the ratio estimator of a population total. Here the estimator is

$$\hat{Y}_R = \hat{R}X, \quad (1)$$

for

$$\hat{R} = \hat{Y}/\hat{X}, \quad \text{and} \quad \hat{Y} = \sum_{k \in s} w_k y_k,$$

where,  $s$  is the set of indices corresponding to sampled units and  $w_k$  is the sampling weight, normalized so that  $\sum w_k$  is an estimator of the population total; *e.g.*,  $w_k = 1/\pi_k$ , where  $\pi_k$  is the first order inclusion probability. The definition of  $\hat{X}$  is analogous to that of  $\hat{Y}$ . Applying total differentials to both sides of (1), we obtain

$$(d\hat{Y}_R) = (d\hat{R})X, \quad (2a)$$

where

$$\begin{aligned} (d\hat{R}) &= \frac{(d\hat{Y})}{\hat{X}} - \frac{\hat{Y}}{\hat{X}^2} (d\hat{X}) \\ &= \frac{1}{\hat{X}} [(d\hat{Y}) - \hat{R}(d\hat{X})]. \end{aligned} \quad (2b)$$

We note that, in general, the total differential for  $\hat{T} = g(\hat{Y}_1, \dots, \hat{Y}_m)$  is given by

$$(d\hat{T}) = \sum \left[ \frac{\partial g(\hat{Y})}{\partial \hat{Y}_i} \right] (d\hat{Y}_i).$$

Although we could have avoided using  $\hat{R}$  in (1) by simply defining

$$Y_{\hat{R}} = \frac{\hat{Y}}{\hat{X}} X,$$

thus removing the need for explicitly defining  $(d\hat{R})$  in (2b), we did so to make the more complex examples, to be given in Section 1.2, clearer. We also note that (2a) does not include the total differential of  $X$ , the population total of the  $x$ -variable, since  $X$  is assumed to be fixed and known.

The next step is to replace all total differentials of estimated quantities by deviations from their respective expected values. On the right hand side, we substitute for  $(d\hat{Y})$  the expression  $(\sum w_k y_k - Y)$ , and so on. For the quantity of interest,  $\hat{Y}_R$ , we replace  $d\hat{Y}_R$  by  $\hat{Y}_R - Y$ . From (2), performing this step, yields

$$\hat{Y}_R - Y \doteq \frac{X}{\hat{X}} \left[ \left( \sum w_k y_k - Y \right) - \hat{R} \left( \sum w_k x_k - X \right) \right]. \quad (3)$$

<sup>1</sup> David A. Binder, Director, Business Survey Methods Division, Statistics Canada, R.H. Coats Building, 11 "A", Ottawa, Ontario, Canada, K1A 0T6.

We see that this expression contains a number of weighted estimators – those that explicitly show their dependence on the  $w_k$ 's, ( $\sum w_k y_k$  and  $\sum w_k x_k$ ) and those where the  $w_k$ 's are implicit in the expression ( $\hat{X}$  and  $\hat{R}$ ).

For the last step, we isolate  $z_k$ , defined by rewriting (3) as

$$\hat{Y}_R - Y \doteq \sum w_k z_k + \text{other terms not depending explicitly on } w_k.$$

Here, we obtain

$$z_k = \frac{X}{\hat{X}} (y_k - \hat{R}x_k). \quad (4)$$

The justification for ignoring the terms not depending explicitly on  $w_k$  will be given in Section 4. Note that  $\sum w_k z_k$  has the form of the estimate of the population total of the variable  $z$ .

Now to obtain the variance of  $\hat{Y}_R$ , we insert the new variable  $z_k$  into the  $k$ -th sample record, and use a standard procedure for estimating the variance of a total, applied to this variable. It is assumed that a variance estimator with good properties is available for the sample design under consideration.

A summary of the method in general is the following:

1. We let the estimator of  $T$  be  $\hat{T}$  and take its total differential. We assume that  $\hat{T}$  is asymptotically design consistent.
2. We replace total differential of  $\hat{T}$ ,  $d\hat{T}$ , by  $\hat{T} - T$ . We replace all other total differentials of estimated quantities by the deviation from their respective expected values, where we substitute for  $(d\hat{Y})$  the expression  $(\sum w_k y_k - Y)$ , and so on.
3. The last step is to isolate  $z_k$ , when we rewrite the result of Step 2 as
 
$$\hat{T} - T \doteq \sum w_k z_k + \text{other terms not depending explicitly on } w_k.$$
4. Finally, to obtain the estimated variance of  $\hat{T}$ , we insert the new variable  $z_k$  into each sampled record, and use the standard procedure (known to have good properties) for estimating the variance of a total, applied to this variable.

### 1.1 Simplest General Case

For one-phase samples, a simple general case is where the estimator can be expressed as a differentiable function of the estimated totals for certain survey variables, some of which may be derived variables at the final sampling unit level. In this case our approach gives:

$$\hat{T} = g(\hat{Y}_1, \dots, \hat{Y}_m)$$

$$(d\hat{T}) = \sum \left[ \frac{\partial g(\hat{Y})}{\partial \hat{Y}_i} \right] (d\hat{Y}_i)$$

$$\begin{aligned} \hat{T} - T &\doteq \sum_i \left[ \frac{\partial g(\hat{Y})}{\partial \hat{Y}_i} \right] \left( \sum_k w_k y_{ik} - Y_i \right) \\ &= \sum w_k z_k + \dots, \end{aligned} \quad (5)$$

where

$$z_k = \sum_i \left[ \frac{\partial g(\hat{Y})}{\partial \hat{Y}_i} \right] y_{ik} = \left[ \frac{\partial g(\hat{Y})}{\partial \hat{Y}} \right]' y_k. \quad (6)$$

In what way is this formulation different from standard Taylor methods? The main difference is how expression (5) is treated. In standard methods, the partial derivatives are evaluated at their expected values before  $z_k$  is derived. Then, for those components of  $z_k$  that are unknown, an estimator is substituted. For the ratio estimator, (1), this would result in  $X/\hat{X}$  disappearing from  $z_k$  in (4), since when  $\hat{X}$  is replaced by its expected value,  $X/\hat{X}$  becomes unity. The  $\hat{R}$  remains in the expression, as it is used to estimate  $R$ , which is needed in the usual derivation of  $z_k$ .

Kott (1990) argues that the variance estimator for the ratio which we have derived has good conditional properties compared to the estimator which leaves out the factor  $X/\hat{X}$ . A number of others have come to similar conclusions. Rao (1995) showed that the method agrees with that obtained from the linearized jackknife. Our conjecture is that since the partial derivatives in expression (5) are evaluated at  $\hat{Y}$  rather than  $Y$ , the linearization is “closer” to the original statistic,  $\hat{T}$ , so that the resulting variances have better properties. This is, of course, not a technical statement, but rather an intuitive justification of the method.

We note that in expression (6) for  $z_k$ , all the terms are directly observed from the sample, so that no substitution of estimators for unknown quantities is needed.

### 1.2 The Case with Extra Parameters

For many examples, the estimator is most easily defined in terms that include the use of parameters that are only used to simplify the definition of the parameter of interest. For the ratio estimator,  $\hat{R}$  is an example of such an *extra parameter*. In this case, an explicit equation for the estimator of the extra parameter is available. The general method in the presence of extra parameters may be written as:

$$\hat{T} = g_1(\hat{Y}_1, \dots, \hat{Y}_m, \hat{\lambda}), \quad \text{where } \hat{\lambda} = g_2(\hat{Y}_1, \dots, \hat{Y}_m),$$

$$(d\hat{T}) = \sum \left[ \frac{\partial g_1(\hat{Y}, \hat{\lambda})}{\partial \hat{Y}_i} \right] (d\hat{Y}_i) + \sum \left[ \frac{\partial g_1(\hat{Y}, \hat{\lambda})}{\partial \hat{\lambda}_j} \right] (d\hat{\lambda}_j),$$



where

$$(d\hat{\lambda}_j) = \sum_i \left[ \frac{\partial g_{2j}(\hat{Y})}{\partial \hat{Y}_i} \right] (d\hat{Y}_i),$$

$$\begin{aligned} \hat{T} - T &\doteq \sum \frac{\partial g_1(\hat{Y}, \hat{\lambda})}{\partial \hat{Y}_i} \left( \sum_k w_k y_{ik} - Y_i \right) \\ &+ \sum \frac{\partial g_1(\hat{Y}, \hat{\lambda})}{\partial \hat{\lambda}_j} \sum_i \frac{\partial g_{2j}(\hat{Y})}{\partial \hat{Y}_i} \left( \sum_k w_k y_{ik} - Y_i \right) \\ &= \sum w_k z_k + \dots, \end{aligned}$$

where

$$z_k = \left[ \frac{\partial g_1(\hat{Y}, \hat{\lambda})}{\partial \hat{Y}} \right]' y_k + \left[ \frac{\partial g_1(\hat{Y}, \hat{\lambda})}{\partial \hat{\lambda}} \right]' \left[ \frac{\partial g_2(\hat{Y})}{\partial \hat{Y}} \right] y_k. \quad (7)$$

For the case where the extra parameters are defined only implicitly through estimating equations, we have the following generalization:

$$\hat{T} = g(\hat{Y}_1, \dots, \hat{Y}_m, \hat{\lambda}),$$

where

$$\hat{U}(\hat{Y}_1, \dots, \hat{Y}_m, \hat{\lambda}) = 0. \quad (8)$$

$$(d\hat{T}) = \sum \left[ \frac{\partial g(\hat{Y}, \hat{\lambda})}{\partial \hat{Y}_i} \right] (d\hat{Y}_i) + \left[ \frac{\partial g(\hat{Y}, \hat{\lambda})}{\partial \hat{\lambda}} \right]' (d\hat{\lambda}),$$

where by taking the total differential of (8) and isolating  $(d\hat{\lambda})$ , we have

$$(d\hat{\lambda}) = - \left[ \frac{\partial \hat{U}(\hat{Y}, \hat{\lambda})}{\partial \hat{\lambda}} \right]^{-1} \sum \left[ \frac{\partial \hat{U}(\hat{Y}, \hat{\lambda})}{\partial \hat{Y}_i} \right] (d\hat{Y}_i). \quad (9)$$

$$\begin{aligned} \hat{T} - T &\doteq \sum_i \left( \frac{\partial g}{\partial \hat{Y}_i} \right) \left( \sum_k w_k y_{ik} - Y_i \right) \\ &- \left( \frac{\partial g}{\partial \hat{\lambda}} \right)' \left[ \frac{\partial \hat{U}}{\partial \hat{\lambda}} \right]^{-1} \sum_i \left( \frac{\partial \hat{U}}{\partial \hat{Y}_i} \right) \left( \sum_k w_k y_{ik} - Y_i \right) \\ &= \sum w_k z_k + \dots, \end{aligned}$$

where

$$z_k = \left[ \frac{\partial g}{\partial \hat{Y}} \right]' y_k - \left[ \frac{\partial g}{\partial \hat{\lambda}} \right]' \left[ \frac{\partial \hat{U}}{\partial \hat{\lambda}} \right]^{-1} \left[ \frac{\partial \hat{U}}{\partial \hat{Y}} \right]' y_k. \quad (10)$$

We see, of course, that (10) is a generalization of the previous forms for  $z_k$  given in (6) and (7).

## 2. OTHER EXAMPLES

Expressions (6), (7) and (10) above are displayed only for the purpose of giving the specific formulae for the various cases. However, in practice, we recommend using the basic steps from first principles. To demonstrate this, we give two examples: one is the familiar Generalized Regression Estimator (GREG); the other gives some new results for the Wilcoxon Rank Sum Test statistic for data from complex surveys.

### 2.1 Generalized Regression Estimator

The usual Generalized Regression Estimator, given, for example, in Särndal, Swensson and Wretman (1989), may be written as

$$\hat{Y}_{GREG} = \hat{Y} + \hat{\beta}'(X - \hat{X}), \quad (11)$$

where the extra parameter  $\hat{\beta}$  is defined as the solution to

$$\sum_k w_k x_k (y_k - x_k' \hat{\beta}) / c_k = 0,$$

where  $c_k$  is the factor to allow for heteroscedastic variance in the regression model. This is equivalent to

$$\hat{S}_{xx} \hat{\beta} - \hat{S}_{xy} = 0, \quad (12)$$

with obvious definitions for  $\hat{S}_{xx}$  and  $\hat{S}_{xy}$ . Taking total differentials in (12) we get

$$(d\hat{S}_{xx})\hat{\beta} + \hat{S}_{xx}(d\hat{\beta}) - (d\hat{S}_{xy}) = 0,$$

so that

$$(d\hat{\beta}) = \hat{S}_{xx}^{-1} [(d\hat{S}_{xy}) - (d\hat{S}_{xx})\hat{\beta}].$$

Therefore, we have

$$\hat{\beta} - \beta \doteq \sum w_k \hat{S}_{xx}^{-1} [x_k (y_k - x_k' \hat{\beta})] / c_k + \dots$$

Now, taking total differentials of (11), we have

$$\begin{aligned} (d\hat{Y}_{GREG}) &= (d\hat{Y}) - \hat{\beta}'(d\hat{X}) + (d\hat{\beta})'(X - \hat{X}) \\ &= (d\hat{Y}) - \hat{\beta}'(d\hat{X}) + \\ &\quad [(\hat{S}_{xy}') - \hat{\beta}'(\hat{S}_{xx}')] \hat{S}_{xx}^{-1} (X - \hat{X}). \end{aligned}$$

After some algebraic manipulation, we obtain

$$\hat{Y}_{GREG} - Y = \sum w_k e_k [1 + x_k' \hat{S}_{xx}^{-1} (X - \hat{X}) / c_k] + \dots,$$

where  $e_k = y_k - x'_k \hat{\beta}$ . We, therefore, define

$$z_k = e_k [1 + x'_k \hat{S}_{xx}^{-1} (X - \hat{X}) / c_k].$$

Taking the variance of the estimated total of this  $z$ -variable is identical to the variance proposed in Särndal, Swensson and Wretman (1989). There, it is argued on the basis of the validity of the regression model, that this variance is preferred to other Taylor expansion estimators for the variance. We see that the derivation of this  $z$ -variable is natural in our approach.

## 2.2 Wilcoxon Rank Sum Statistic

We now show how our method works in the case of a more difficult non-standard case. We assume that our sampled units belong to one of two subpopulations which we name Population 1 and Population 2. We define

$$I\{x \leq y\} = \begin{cases} 1 & \text{if } x \leq y, \\ 0 & \text{otherwise,} \end{cases} \text{ and } \delta_k = \begin{cases} 1 & \text{if } k \in \text{Pop. 1} \\ 0 & \text{otherwise.} \end{cases}$$

We let

$$\hat{N}_1(t) = \sum_{k \in S} w_k \delta_k I\{x_k \leq t\},$$

which corresponds to the estimated number of Population 1 units that have values less than or equal to  $t$ . We define  $\hat{N}_2(t)$  analogously. We denote  $\hat{N}_j = \hat{N}_j(\infty)$ , the estimated number of units in Population  $j$ . Now a weighted version of the Wilcoxon Rank Sum Test statistic is

$$\hat{T}_W = \int_0^\infty [\hat{N}_1(t) + \hat{N}_2(t)] d\hat{N}_1(t). \quad (13)$$

This corresponds to the weighted sum of the ranks from Population 1 among the weighted ranks of the combined sample. To derive the asymptotic expected value of  $\hat{T}_W$  in (13), we let  $N_i(t) = E[\hat{N}_i(t)]$  for  $i = 1, 2$ , and substitute  $N_i(t)$  for  $\hat{N}_i(t)$  in (13). We then define  $F_i(t) = N_i(t)/N_i$ , where  $N_i = E(\hat{N}_i)$  and we give the null hypothesis as  $F_1(t) = F_2(t) = F(t)$ , say. This results in the asymptotic expectation being

$$\int_0^1 (N_1 + N_2) F(t) N_1 dF(t) = N_1 (N_1 + N_2) / 2.$$

Note that in the case of independent samples of size  $N_1$  and  $N_2$  from Population 1 and Population 2, respectively, where each population is assumed to have a continuous distribution function and the samples are taken using simple random sampling, the exact expected value for  $\hat{T}_W$  in (13) is  $N_1 (N_1 + N_2 + 1) / 2$ .

We consider the statistic

$$\hat{T}_W^* = \int_0^\infty [\hat{N}_1(t) + \hat{N}_2(t)] d\hat{N}_1(t) - \frac{\hat{N}_1 (\hat{N}_1 + \hat{N}_2)}{2}.$$

We use  $\Delta$  rather than  $d$  to denote the total differential, since  $d$  is used under the integral. Therefore, we have

$$\begin{aligned} (\Delta \hat{T}_W^*) &= \int_0^\infty [\Delta \hat{N}_1(t) + \Delta \hat{N}_2(t)] d\hat{N}_1(t) \\ &\quad + \int_0^\infty [\hat{N}_1(t) + \hat{N}_2(t)] d\Delta \hat{N}_1(t) \\ &\quad - \frac{(\Delta \hat{N}_1) (\hat{N}_1 + \hat{N}_2) + \hat{N}_1 (\Delta \hat{N}_1 + \Delta \hat{N}_2)}{2}. \end{aligned}$$

Continuing with our usual approach, we have

$$\begin{aligned} \hat{T}_W^* - T_W^* &\doteq \int_0^\infty \left( \sum w_k I\{x_k \leq t\} \right) d\hat{N}_1(t) \\ &\quad + \sum w_k \delta_k [\hat{N}_1(x_k) + \hat{N}_2(x_k)] \\ &\quad - \frac{\sum w_k \delta_k (\hat{N}_1 + \hat{N}_2) + \hat{N}_1 \sum w_k}{2} + \dots, \end{aligned}$$

so that

$$\begin{aligned} z_k &= \sum_j w_j \delta_j I\{x_k \leq x_j\} + \delta_k [\hat{N}_1(x_k) + \hat{N}_2(x_k)] \\ &\quad - \frac{\delta_k (\hat{N}_1 + \hat{N}_2) + \hat{N}_1}{2}. \end{aligned} \quad (14)$$

We are not aware of this result previously being documented. It can be shown that when the null hypothesis is true and we select independently from two populations using simple random sampling, where the populations have continuous distribution functions, the variance we obtain from the  $z$ -variables in (14) is asymptotically equivalent to the usual classical formula.

## 3. TWO-PHASE SAMPLES

The method described above extends quite easily to the case of two-phase samples. For example, consider the two-phase ratio estimator of the population total, given by

$$\hat{Y}_{R(2)} = \frac{\hat{Y}}{\hat{X}} \hat{X}^{(1)} = \hat{R} \hat{X}^{(1)}, \quad (15)$$



where  $\hat{X}^{(1)} = \sum w_k x_k$  is the first phase estimate of  $X$  based on first phase weights  $\{w_k\}$ , and  $\hat{Y}$  and  $\hat{X}$  are the estimates of  $Y$  and  $X$ , respectively, based the second phase sample units with weights  $\{w_k w_{2k}\}$ , where  $w_{2k}$  is the weight assigned to the selected second phase unit, conditional on being in the first phase sample. In particular, letting

$$a_k = \begin{cases} 1 & \text{if the } k\text{-th unit is in the second phase sample,} \\ 0 & \text{otherwise,} \end{cases}$$

we have

$$\hat{Y} = \sum_{k \in s} w_k w_{2k} a_k y_k,$$

where  $s$  is the set of indices corresponding to units in the first phase sample.

Taking total differentials of (15), we have

$$(d\hat{Y}_{R(2)}) = \left( \frac{\hat{X}^{(1)}}{\hat{X}} \right) [(d\hat{Y}) - \hat{R}(d\hat{X})] + \hat{R}(d\hat{X}^{(1)}).$$

We now replace the total differentials by weighted sums over first phase units:

$$\hat{Y}_{R(2)} - \hat{Y} \doteq$$

$$\sum_{k \in s} w_k \left[ a_k w_{2k} \left( \frac{\hat{X}^{(1)}}{\hat{X}} \right) (y_k - \hat{R}x_k) + \hat{R}x_k \right] + \dots,$$

so that

$$z_k = a_k w_{2k} \left( \frac{\hat{X}^{(1)}}{\hat{X}} \right) (y_k - \hat{R}x_k) + \hat{R}x_k. \quad (16)$$

We see that the steps we have taken are essentially the same as in the one phase sample case. However, it is important to note that now  $z_k$  contains the random variable,  $a_k$ , that is used to indicate whether or not the sample unit is in the second phase sample. This is needed to compute the two phase variance estimator.

Variances obtained from the  $z$ -variable in (16) are identical to those given in Rao and Sitter (1995), who used a linearization of the jackknife to obtain their results.

Extensions to other estimation problems in two phase samples are straightforward. Suppose, for example, that  $(\hat{Y}_1, \dots, \hat{Y}_m)$  are estimates of  $(Y_1, \dots, Y_m)$  from the second phase samples, and that  $(\hat{X}_1^{(1)}, \dots, \hat{X}_p^{(1)})$  are estimates of variables available only for first phase sample units. We suppose that a set of extra parameters,  $\lambda$ , are defined only in terms of the units in the second phase, and that the variable of interest is defined in terms of these extra parameters and the  $\hat{X}_j^{(1)}$ 's. Formally, then, we have

$$U(\hat{\lambda}, \hat{Y}) = 0,$$

and

$$\hat{T} = g(\hat{X}^{(1)}, \hat{\lambda}).$$

Taking total differentials, we have as in (9),

$$(d\hat{\lambda}) = - \left[ \frac{\partial \hat{U}}{\partial \hat{\lambda}} \right]^{-1} \left[ \frac{\partial \hat{U}}{\partial \hat{Y}} \right] (d\hat{Y}),$$

so that

$$\begin{aligned} \hat{T} - T &\doteq \left[ \frac{\partial g}{\partial \hat{X}^{(1)}} \right]' \left( \sum_k w_k x_k - X \right) \\ &\quad - \left[ \frac{\partial g}{\partial \hat{\lambda}} \right]' \left[ \frac{\partial \hat{U}}{\partial \hat{\lambda}} \right]^{-1} \left[ \frac{\partial \hat{U}}{\partial \hat{Y}} \right] \left( \sum_k a_k w_k w_{2k} y_k - Y \right). \end{aligned}$$

Therefore, the general expression for  $z_k$  is

$$z_k = \left[ \frac{\partial g}{\partial \hat{X}^{(1)}} \right]' x_k - \left[ \frac{\partial g}{\partial \hat{\lambda}} \right]' \left[ \frac{\partial \hat{U}}{\partial \hat{\lambda}} \right]^{-1} \left[ \frac{\partial \hat{U}}{\partial \hat{Y}} \right] a_k w_{2k} y_k.$$

It then becomes necessary to put the  $z$ -variable into the algorithm that estimates the variance of the estimator of a total from a two phase sample.

#### 4. JUSTIFICATION

The technique we have described can be considered as a direct result of the formulation given in Binder and Patak (1994). We will summarize one of the main results in that paper. Suppose we are interested in parameter  $\theta$ , defined as the solution to

$$\hat{U}_1(\theta, \hat{\lambda}_\theta) = \sum_{k \in s} w_k u_1(y_k, \theta, \hat{\lambda}_\theta) = 0,$$

where  $\hat{\lambda}_\theta$  is the estimate of an extra parameter, defined as the solution to

$$\hat{U}_2(\theta, \hat{\lambda}_\theta) = \sum_{k \in s} w_k u_2(y_k, \theta, \hat{\lambda}_\theta) = 0,$$

for a given  $\theta$ . Through an argument based on removing extra parameters for problems of testing hypotheses on  $\theta$ , Binder and Patak recommend basing inferences about  $\theta$  on the variable

$$u^* = u_1(y, \theta, \hat{\lambda}_\theta) - \left[ \frac{\partial \hat{U}_1}{\partial \hat{\lambda}_\theta} \right] \left[ \frac{\partial \hat{U}_2}{\partial \hat{\lambda}_\theta} \right]^{-1} u_2(y, \theta, \hat{\lambda}_\theta). \quad (17)$$

In particular, two-sided confidence intervals for  $\theta$  are to be based on

$$\left\{ \theta \mid \frac{\hat{U}_1^2(\theta, \hat{\lambda}_\theta)}{\hat{W}} \leq \chi_{1-\alpha}^2(1) \right\},$$

where  $\hat{W}$  is the estimated variance of the estimator of a total when the variable being estimated is  $u^*$ .

We let  $u_1 = g(\lambda_1, \lambda_2) - \theta$ . The kernel of the estimating equations for the  $y$ -totals will be given by  $u_{21} = y - \lambda_1$  and the kernel of the estimating equations for  $\lambda_2$  is given by  $u_{22}(\lambda_1, \lambda_2)$ . We let

$$\hat{U}_2 = \sum w_k \begin{bmatrix} u_{21} \\ u_{22} \end{bmatrix} = \begin{bmatrix} \hat{Y} - \hat{N}\hat{\lambda}_1 \\ \hat{N}u_{22} \end{bmatrix}, \text{ where } \hat{N} = \sum w_k.$$

After some algebra, from (17) the variance of interest is the variance of the estimated total based on the variable  $u^*$ , given by,

$$\begin{aligned} & \left[ \frac{\partial g(\hat{\lambda}_1, \hat{\lambda}_2)}{\partial \hat{\lambda}_1} \right]'_y \\ & - \left[ \frac{\partial g(\hat{\lambda}_1, \hat{\lambda}_2)}{\partial \hat{\lambda}_2} \right]' \left[ \frac{\partial u_{22}(\hat{\lambda}_1, \hat{\lambda}_2)}{\partial \hat{\lambda}_2} \right]^{-1} \left[ \frac{\partial u_{22}(\hat{\lambda}_1, \hat{\lambda}_2)}{\partial \hat{\lambda}_1} \right]_y \\ & + \text{constant terms.} \end{aligned}$$

This is equivalent to expression (10), thus showing that the methods here are consistent with those in Binder and Patak (1994).

## ACKNOWLEDGEMENTS

I wish to thank Georgia Roberts and Alain Théberge who had many useful suggestions on improving the readability of an earlier draft. I am also grateful to J.N.K. Rao for many useful discussions on this topic, and to an anonymous referee for some constructive remarks.

## REFERENCES

- BINDER, D.A., and PATAK, Z. (1994). Use of estimating functions for interval estimation from complex surveys. *Journal of the American Statistical Association*, 89, 1035-1043.
- KOTT, P.S. (1990). Estimating the conditional variance of a design consistent regression estimator. *Journal of Statistical Planning and Inference*, 24, 287-296.
- RAO, J.N.K. (1995). Private communication.
- RAO, J.N.K., and SITTER, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82, 453-460.
- SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J.H. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76, 527-537.



# Jackknife Linearization Variance Estimators Under Stratified Multi-Stage Sampling

W. YUNG and J.N.K. RAO<sup>1</sup>

## ABSTRACT

Variance estimation for the poststratified estimator and the generalized regression estimator of a total under stratified multi-stage sampling is considered. By linearizing the jackknife variance estimator, a jackknife linearization variance estimator is obtained which is different from the standard linearization variance estimator. This variance estimator is computationally simpler than the jackknife variance estimator and yet leads to values close to the jackknife. Properties of the jackknife linearization variance estimator, the standard linearized variance estimator, and the jackknife variance estimator are studied through a simulation study. All of the variance estimators performed well both unconditionally and conditionally given a measure of how far away the estimated totals of auxiliary variables are from the known population totals. A jackknife variance estimator based on incorrect reweighting performed poorly, indicating the importance of correct reweighting when using the jackknife method.

**KEY WORDS:** Generalized regression estimator; Jackknife variance estimator; Linearized variance estimator; Poststratified estimator.

## 1. INTRODUCTION

Large-scale sample surveys often use stratified multi-stage designs with large numbers of strata,  $L$ , and relatively few primary sampling units (clusters),  $n_h (\geq 2)$ , sampled within each stratum. Within each cluster, some elements (ultimate units) are sampled according to some sampling method. We do not specify the number of stages or the sampling methods used after the first-stage sampling, but we assume that subsampling within sampled clusters is performed to ensure unbiased estimation of cluster totals,  $Y_{hi}$ ,  $i = 1, \dots, n_h$ ;  $h = 1, \dots, L$ .

From the specification of the survey design, basic weights  $w_{hik} (> 0)$ , attached to the  $(hik)$ -th element, are obtained. Often these basic weights  $w_{hik}$  are subjected to poststratification adjustment to ensure consistency with known totals of poststratification variables. In the case of a single poststratifier, the weights are ratio-adjusted to the known population counts (e.g., age-sex counts). To handle two or more poststratifiers with known marginal population counts, the weights  $w_{hik}$  can be calibrated through generalized regression (see section 4), as in the Canadian Labour Force Survey (CLFS).

The CLFS uses the jackknife method for estimating the variance of the generalized regression estimator. The jackknife method is computer intensive but it is readily applicable to general smooth statistics, unlike the linearization method. Moreover, it possesses good conditional properties. For example, in the context of simple random sampling and the ratio estimator, Royall and Cumberland (1981) showed that the jackknife variance estimator tracks the conditional variance given the sample mean of the auxiliary variable  $x$ .

The main purpose of this paper is to study variance estimation for the ratio-adjusted poststratified estimator and the generalized regression estimator under stratified sampling. By linearizing the jackknife variance estimator, a jackknife linearization variance estimator is obtained which is different from the standard linearization variance estimator. In the case of the poststratified estimator, this variance estimator is identical to Rao's (1985) variance estimator. The proposed variance estimator is computationally simpler than the jackknife variance estimator and yet leads to values close to the jackknife.

Section 2 introduces the jackknife variance estimator for the basic expansion estimator of the total,  $Y$ . Section 3 presents the jackknife and the jackknife linearization variance estimators for the poststratified estimator. These results are extended in section 4 to the generalized regression estimator in the context of multiple poststratification variables. Section 5 deals with variance estimation for a ratio of two totals, both of which are estimated using a generalized regression estimator. Results of a simulation study on the relative performances of the usual linearization variance estimator, the jackknife and the jackknife linearization variance estimators are reported in section 6.

## 2. BASIC ESTIMATOR

Using the basic weights  $w_{hik}$ , an unbiased estimator of the population total  $Y$  is of the form

$$\hat{Y} = \sum_{(hik) \in S} w_{hik} y_{hik}, \quad (2.1)$$

<sup>1</sup> W. Yung, Statistics Canada, Household Survey Methods Division, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, K1A 0T6; and J.N.K. Rao, Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, K1S 5B6.

where  $s$  denotes the sample of elements and  $y_{hik}$  is the value of the characteristic of interest associated with the sample element  $(hik) \in s$ . For simplicity, we assume complete response in this paper.

It is common practice to sample clusters without replacement. However, at the stage of variance estimation, the calculations are greatly simplified by treating the sample as if the clusters are sampled with replacement. This approximation generally leads to overestimation of the variance of  $\hat{Y}$ , but the relative bias is likely to be small if the first-stage sampling fractions are small.

An estimator of the variance of  $\hat{Y}$  is given by

$$v(\hat{Y}) = \sum_{h=1}^L \frac{1}{n_h(n_h - 1)} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2 = v(y_{hi}), \quad (2.2)$$

where  $y_{hi} = \sum_k (n_h w_{hik}) y_{hik}$ , and  $\bar{y}_h = (1/n_h) \sum_i y_{hi}$ . The operator notation  $v(y_{hi})$  denotes that  $v(\hat{Y})$  depends only on the  $y_{hi}$ 's.

To introduce the jackknife method, we need the estimator  $\hat{Y}_{(gj)}$  for each  $(gj)$  obtained from the sample after omitting the data from the  $j$ -th sampled cluster in the  $g$ -th stratum ( $j = 1, \dots, n_g$ ;  $g = 1, \dots, L$ ). It is simply obtained from (2.1) by letting  $w_{gik} = 0$ , changing  $w_{gik}$  ( $i \neq j$ ) to  $n_g w_{gik} / (n_g - 1)$  and retaining the original weights  $w_{hik}$  for  $h \neq g$ , i.e.,

$$w_{hik(gj)} = \begin{cases} 0 & \text{if } (hi) = (gj) \\ \frac{n_g}{(n_g - 1)} w_{gik} & \text{if } h = g \text{ and } i \neq j \\ w_{hik} & \text{if } h \neq g. \end{cases}$$

These jackknife weights,  $w_{hik(gj)}$ , are calculated for each cluster  $(gj)$ . The resulting estimator of  $Y$  is

$$\hat{Y}_{(gj)} = \sum_{(hik) \in s} w_{hik(gj)} y_{hik}.$$

The jackknife variance estimator is then given by

$$v_J(\hat{Y}) = \sum_{g=1}^L \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\hat{Y}_{(gj)} - \hat{Y})^2. \quad (2.3)$$

The variance estimator (2.3) is applicable to general smooth statistics, say  $\hat{\theta} = g(\hat{Y})$ , by simply replacing  $\hat{Y}_{(gj)}$  and  $\hat{Y}$  with  $\hat{\theta}_{(gj)} = g(\hat{Y}_{(gj)})$  and  $\hat{\theta}$  respectively. In the linear case,  $\hat{\theta} = \hat{Y}$ , the jackknife variance estimator is identical to the customary variance estimator (2.2).

### 3. POSTSTRATIFIED ESTIMATOR

Suppose the population is partitioned into  $C$  poststrata with known population counts  ${}_cM$ ,  $c = 1, \dots, C$ . We will use the prescript  $c$  to denote poststrata. An estimator of  ${}_cM$  is given by

$${}_c\hat{M} = \sum_{(hik) \in {}_cS} w_{hik}, \quad (3.1)$$

where  ${}_cS$  is the sample of elements belonging to the  $c$ -th poststratum. Similarly, an estimator of the poststratum total  ${}_cY$  is

$${}_c\hat{Y} = \sum_{(hik) \in {}_cS} w_{hik} y_{hik}.$$

Using the estimators  ${}_c\hat{Y}$  and  ${}_c\hat{M}$ , we obtain a poststratified estimator of the total  $Y$  as

$$\hat{Y}_{ps} = \sum_c \frac{{}_cM}{{}_c\hat{M}} {}_c\hat{Y}. \quad (3.2)$$

We can rewrite (3.2) as

$$\hat{Y}_{ps} = \sum_c \sum_{(hik) \in {}_cS} {}_c w_{hik} y_{hik}$$

where  ${}_c w_{hik} = w_{hik} ({}_cM / {}_c\hat{M})$  is the ratio-adjusted weight for  $(hik) \in {}_cS$ . If  $y_{hik}$  is the indicator variable for a poststratum, say  $c$ , then  $\hat{Y}_{ps} = {}_c\hat{M}$ , thus ensuring consistency with known totals,  ${}_cM$ .

The standard linearization variance estimator is given by (2.2) with  $y_{hi}$  changed to

$$\tilde{e}_{hi} = \sum_c \sum_{k \in {}_cS} (n_h w_{hik}) {}_c e_{hik},$$

where  ${}_c e_{hik} = y_{hik} - {}_c\hat{Y} / {}_c\hat{M}$  for the  $k$ -th element in the  $(hi)$ -th cluster belonging to  ${}_cS$ , i.e.,

$$v_L(\hat{Y}_{ps}) = v(\tilde{e}_{hi}). \quad (3.3)$$

Rao (1985) proposed an alternative linearization variance estimator using the ratio-adjusted weights  ${}_c w_{hik}$ :

$$v_R(\hat{Y}_{ps}) = v(e_{hi}^*) \quad (3.4)$$

where

$$e_{hi}^* = \sum_c \sum_{k \in {}_cS} (n_h {}_c w_{hik}) {}_c e_{hik}.$$

Turning to the jackknife method, we need to recalculate the poststratification weights  ${}_c w_{hik}$  each time a cluster  $(gj)$  is deleted. This is done by using the jackknife weights  $w_{hik(gj)}$  in (3.1) to get  ${}_c\hat{M}_{(gj)}$  and then using  ${}_c w_{hik(gj)} = ({}_cM / {}_c\hat{M}_{(gj)}) w_{hik(gj)}$  to get



$$\hat{Y}_{ps(gj)} = \sum_c \sum_{(hik) \in cs} {}_c w_{hik(gj)} y_{hik}.$$

The jackknife variance estimator is then obtained as

$$v_J(\hat{Y}_{ps}) = \sum_{g=1}^L \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\hat{Y}_{ps(gj)} - \hat{Y}_{ps})^2. \quad (3.5)$$

By linearizing (3.5), we obtain a jackknife linearization variance estimator,  $v_{JL}(\hat{Y}_{ps})$ , which is identical to Rao's variance estimator (3.4); see also Valliant (1993). In the important special case of  $n_h = 2$  clusters per stratum, (3.4) and (3.5) are in fact asymptotically equal to higher order terms, as the number of strata  $L$  increases (Yung 1996).

Rao (1985) justified (3.4) on heuristic grounds by noting that for simple random sampling it reduces to a conditionally valid variance estimator given the poststrata sample sizes, unlike the standard linearization variance estimator (3.3). Särndal, Swensson and Wretman (1989) obtained a variance estimator of the form (3.4) in the context of unistage sampling under a model-assisted framework. Since  $v_{JL}(\hat{Y}_{ps})$  and  $v_J(\hat{Y}_{ps})$  are approximately equal, the foregoing results suggest that both variance estimators should be "robust" in the sense of possessing good conditional properties given the estimated poststrata counts. Valliant (1993) conducted a simulation study to demonstrate the "robustness" of  $v_J(\hat{Y}_{ps})$  and  $v_{JL}(\hat{Y}_{ps})$ .

#### 4. GENERALIZED REGRESSION ESTIMATOR

In practice, it is common to form poststrata according to two or more auxiliary variables. If the resulting cell level population counts are available, the ratio-adjusted poststratified estimator can be used to increase the efficiency of the estimates. However, these cell counts may not be known in practice. For instance, marginal counts may be known only for age groups and race groups but not cell counts for the individual age-race groups. This means that in terms of a two-way table, the marginal counts are known but not the cell level counts. To handle several poststratifiers with known marginal population counts, we can use a generalized regression estimator of  $Y$  by using indicator auxiliary variables to denote the categories of the poststratifiers (Huang and Fuller 1978; Deville and Särndal 1992).

Let  $\mathbf{x}_{hik}$  be a vector of auxiliary variables with known population totals  $\mathbf{X}$ . The generalized regression estimator of  $Y$  is then given by

$$\hat{Y}_r = \hat{Y} + (\mathbf{X} - \hat{\mathbf{X}})^T \hat{\mathbf{B}}, \quad (4.1)$$

where

$$\hat{\mathbf{X}} = \sum_{(hik) \in s} w_{hik} \mathbf{x}_{hik},$$

and  $\hat{\mathbf{B}}$  is the vector of estimated regression coefficients

$$\hat{\mathbf{B}} = \hat{\mathbf{A}}^{-1} \hat{\mathbf{b}},$$

where

$$\hat{\mathbf{A}} = \sum_{(hik) \in s} w_{hik} \mathbf{x}_{hik} \mathbf{x}_{hik}^T,$$

and

$$\hat{\mathbf{b}} = \sum_{(hik) \in s} w_{hik} \mathbf{x}_{hik} y_{hik}.$$

The poststratified estimator,  $\hat{Y}_{ps}$ , is a special case of (4.1) by letting  $\mathbf{x}_{hik}$  denote the vector of indicator variables for the poststrata. In this case,  $\hat{\mathbf{X}} = ({}_1\hat{M}, \dots, {}_c\hat{M})^T$ ,  $\mathbf{X} = ({}_1M, \dots, {}_cM)^T$ , and  $\hat{\mathbf{B}} = ({}_1\hat{R}, \dots, {}_c\hat{R})^T$  with  ${}_c\hat{R} = {}_c\hat{Y}/{}_c\hat{M}$ . Thus,

$$\hat{Y}_r = \hat{Y} + \sum_c {}_c\hat{R}({}_cM - {}_c\hat{M}) = \hat{Y}_{ps}.$$

In the case of two or more poststratifiers,  $\mathbf{X}$  corresponds to the vector of marginal population counts.

The generalized regression estimator may be rewritten as

$$\hat{Y}_r = \sum_{(hik) \in s} w_{hik}^* y_{hik},$$

where

$$w_{hik}^* = w_{hik} a_{hik} \quad (4.2)$$

is the "final" or "calibration" weight with

$$a_{hik} = 1 + \mathbf{x}_{hik}^T \hat{\mathbf{A}}^{-1} (\mathbf{X} - \hat{\mathbf{X}}).$$

In the special case of  $\hat{Y}_{ps}$ , we have  $a_{hik} = {}_cM/{}_c\hat{M}$  for  $(hik) \in cs$ . Writing  $\hat{Y}_r$  in the operator notation as  $\hat{Y}_r(y_{hik})$ , it is readily verified that the generalized regression estimator  $\hat{\mathbf{X}}_r = \hat{Y}_r(\mathbf{x}_{hik}) = \mathbf{X}$ , thus ensuring consistency with known totals  $\mathbf{X}$ .

Turning to variance estimation, the standard linearization variance estimator is again given by (2.2) with  $y_{hi}$  changed to

$$\tilde{e}_{hi} = \sum_k (n_h w_{hik}) e_{hik},$$

where

$$e_{hik} = y_{hik} - \mathbf{x}_{hik}^T \hat{\mathbf{B}} \quad (4.3)$$

are the estimated residuals, i.e.,

$$v_L(\hat{Y}_r) = v(\tilde{e}_{hi}). \quad (4.4)$$

For the jackknife method we need to recalculate the calibration weights  $w_{hik}^*$  each time a cluster ( $gj$ ) is deleted. These weights are given by

$$w_{hik}^*(gj) = w_{hik}(gj) a_{hik}(gj),$$

where

$$a_{hik}(gj) = 1 + \mathbf{x}_{hik}^T \hat{\mathbf{A}}_{(gj)}^{-1} (\mathbf{X} - \hat{\mathbf{X}}_{(gj)}),$$

$$\hat{\mathbf{A}}_{(gj)} = \sum_{(hik) \in s} w_{hik}(gj) \mathbf{x}_{hik} \mathbf{x}_{hik}^T,$$

and

$$\hat{\mathbf{X}}_{(gj)} = \sum_{(hik) \in s} w_{hik}(gj) \mathbf{x}_{hik}.$$

Denote the resulting generalized regression estimator as

$$\begin{aligned} \hat{Y}_{r(gj)} &= \sum_{(hik) \in s} w_{hik}^*(gj) y_{hik} \\ &= \hat{Y}_{(gj)} + (\mathbf{X} - \hat{\mathbf{X}}_{(gj)})^T \hat{\mathbf{B}}_{(gj)} \end{aligned}$$

where  $\hat{\mathbf{B}}_{(gj)}$  is the vector of estimated regression coefficients when the ( $gj$ )-th cluster is deleted:

$$\hat{\mathbf{B}}_{(gj)} = \hat{\mathbf{A}}_{(gj)}^{-1} \hat{\mathbf{b}}_{(gj)}$$

with

$$\hat{\mathbf{b}}_{(gj)} = \sum_{(hik) \in s} w_{hik}(gj) \mathbf{x}_{hik} y_{hik}.$$

The jackknife variance estimator of  $\hat{Y}_r$  is then given by

$$v_J(\hat{Y}_r) = \sum_{g=1}^L \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\hat{Y}_{r(gj)} - \hat{Y}_r)^2. \quad (4.5)$$

It is shown in the Appendix that by linearizing the jackknife variance estimator (4.5), one obtains

$$v_{JL}(\hat{Y}_r) = v(e_{hi}^*) \quad (4.6)$$

with

$$e_{hi}^* = \sum_k (n_h w_{hik}^*) e_{hik}$$

where  $w_{hik}^*$  is defined in (4.2) and  $e_{hik}$  is defined in (4.3). It is interesting to note that the jackknife linearization variance estimator (4.6) is similar to the model-assisted variance estimator proposed by Särndal, Swensson and Wretman (1989) in the context of unistage sampling. Yung (1996) established the asymptotic equivalence of  $v_J(\hat{Y}_r)$  and  $v_{JL}(\hat{Y}_r)$  to higher order terms in the important special case of  $n_h = 2$  clusters per stratum. Note that the above results are also applicable to general auxiliary variables,  $\mathbf{x}_{hik}$ .

Binder (1996) proposed a new linearization method which also leads to  $v_{JL}(\hat{Y}_r)$ . In this method, the partial derivatives are evaluated at the estimates  $\hat{Y}$ ,  $\hat{\mathbf{X}}$  and  $\hat{\mathbf{B}}$ , rather than the population values  $Y$ ,  $\mathbf{X}$  and  $\mathbf{B}$  as in the traditional linearization method. Given that  $v_J$  and  $v_{JL}$  are design-consistent (Yung 1996) and possess good conditional properties, our results provide theoretical justification for Binder's method which was proposed as a "cookbook approach".

The computation of the jackknife variance estimator involves the inversion of the matrix  $\hat{\mathbf{A}}_{(gj)}$  for each ( $gj$ ). However, the jackknife variance estimator can be approximated by retaining the inverse for the full sample,  $\hat{\mathbf{A}}^{-1}$ , and then using modified weights

$$\tilde{w}_{hik}(gj) = w_{hik}(gj) \tilde{a}_{hik}(gj)$$

with

$$\tilde{a}_{hik}(gj) = 1 + (w_{hik}/w_{hik}(gj)) \mathbf{x}_{hik}^T \hat{\mathbf{A}}^{-1} (\mathbf{X} - \hat{\mathbf{X}}_{(gj)}).$$

The resulting estimator of  $Y$ , when the ( $gj$ )-th cluster is deleted, is given by

$$\tilde{Y}_{r(gj)} = \sum_{(hik) \in s} \tilde{w}_{hik}(gj) y_{hik}$$

and the corresponding jackknife variance estimator is

$$v_{J1}(\hat{Y}_r) = \sum_{g=1}^L \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\tilde{Y}_{r(gj)} - \hat{Y}_r)^2. \quad (4.7)$$

It is readily seen that (4.7) is exactly equal to the standard linearization variance estimator (4.4).

## 5. ESTIMATION OF A RATIO

Often a ratio of two estimated totals is required. For example, in a family expenditure survey, one may be interested in the proportion of income spent on clothing. Let

$$\hat{Y}_r = \hat{Y} + (\mathbf{X} - \hat{\mathbf{X}})^T \hat{\mathbf{B}}_1$$

be a generalized regression estimator of the total amount spent on clothing,  $Y$ . Similarly, let

$$\hat{Z}_r = \hat{Z} + (\mathbf{X} - \hat{\mathbf{X}})^T \hat{\mathbf{B}}_2$$

be a generalized regression estimator of the total income,  $Z$ . The proportion of interest is  $\theta = Y/Z$ , and can be estimated by

$$\hat{\theta} = \hat{Y}_r / \hat{Z}_r.$$



The jackknife variance estimator is given by

$$v_J(\hat{\theta}) = \sum_g \frac{n_g - 1}{n_g} \sum_j (\hat{\theta}_{(gj)} - \hat{\theta})^2 \quad (5.1)$$

where

$$\hat{\theta}_{(gj)} = \hat{Y}_{r(gj)} / \hat{Z}_{r(gj)}.$$

Linearizing the jackknife variance estimator, (5.1), we obtain a jackknife linearization variance estimator

$$v_{JL}(\hat{\theta}) = v(r_{hi}^{**}) \quad (5.2)$$

where

$$r_{hi}^{**} = \frac{1}{\hat{Z}_r} \sum_k (n_h w_{hik}^*) e_{hik}^*$$

with

$$e_{hik}^* = e_{hik} - \frac{\hat{Y}_r}{\hat{Z}_r} \tilde{e}_{hik},$$

and

$$e_{hik} = y_{hik} - \mathbf{x}_{hik}^T \hat{\mathbf{B}}_1, \quad \tilde{e}_{hik} = z_{hik} - \mathbf{x}_{hik}^T \hat{\mathbf{B}}_2.$$

Proof of (5.2) is omitted for simplicity.

## 6. SIMULATION STUDY

We performed a simulation study to investigate the unconditional and conditional finite sample properties of the variance estimators in the case of a single poststratifier as well as two poststratification variables. For this purpose, we used a fixed finite population, considered by Valliant (1993), consisting of 10,841 persons included in the September 1988 Current Population Survey (CPS) of the United States. The variable of interest,  $y$ , is the weekly wages for each person. The single poststratifier was defined on the basis of age, race and sex, while the two poststratifiers were based on the variables age, with five levels, and race, with two levels (see Tables 1 and 2 for details).

Table 1

Assignment of Age/Race/Sex Categories to Poststrata:  
Single Poststratifier

Age	Nonblack		Black	
	Male	Female	Male	Female
19 and under	1	1	1	1
20-24	2	3	3	3
25-34	5	6	4	4
35-64	7	8	4	4
65 and over	2	3	3	1

Note: Cell numbers (1-8) are poststratum identification numbers.

Table 2

Assignment of Age/Race Categories to Poststrata:  
Two Poststratifiers

Age	Nonblack	Black	
19 and under	(1,1)	(1,2)	PS1(1)
20-24	(2,1)	(2,2)	PS1(2)
25-34	(3,1)	(3,2)	PS1(3)
35-64	(4,1)	(4,2)	PS1(4)
65 and over	(5,1)	(5,2)	PS1(5)
	PS2(1)	PS2(2)	

Note: Number in margins are poststratum identification numbers.  
Cells  $(i,j)$  denote poststrata ( $i = 1, \dots, 5; j = 1, 2$ ).

The study population contained 2,826 geographical segments, each composed of about four neighbouring households. One hundred design strata ( $L = 100$ ) were created with each stratum having about the same total number of households. We used a stratified two-stage sampling design with segments as clusters and persons as the second-stage units. In each stratum  $n_h = 2$  segments were selected with probability proportional to the number of persons in each segment, and a simple random sample of  $m_{hi} = 4$  persons was selected without replacement if the sample segment contained more than four persons. In sample segments with four or fewer persons, all persons in the segment were selected. Using this design, we selected two sets of 10,000 independent samples, one set for the one-way poststratification case and the other set for the two-way poststratification case.

From each sample, we computed the basic estimator, the relevant poststratified estimator,  $\hat{Y}_{ps}$  or  $\hat{Y}_r$ , and four variance estimators: the standard linearization variance estimator  $v_L$ , the jackknife linearization variance estimator  $v_{JL}$ , the jackknife  $v_J$ , and an incorrect jackknife variance estimator  $v_J^*$ . In applying the jackknife procedure, it is questioned whether or not the “final” or “calibrated” weights need to be recalculated each time a cluster is deleted. The correct jackknife variance estimator does recalculate the “final” weight whenever a cluster is deleted while the incorrect jackknife variance estimator fails to do this. For the one-way poststratification case,  $v_J^*(\hat{Y}_{ps})$  uses the full adjustment  ${}_cM/{}_c\hat{M}$  instead of  ${}_cM/{}_c\hat{M}_{(gj)}$  when the  $(gj)$ -th cluster is deleted, *i.e.*,  $\hat{Y}_{ps(gj)}$  uses the weights  $({}_cM/{}_c\hat{M})w_{hik(gj)}$  instead of  $({}_cM/{}_c\hat{M}_{(gj)})w_{hik(gj)}$ . Similarly, for the two-way poststratification case,  $v_J^*(\hat{Y}_r)$  uses the full adjustment  $a_{hik}$  instead of  $a_{hik(gj)}$  when the  $(gj)$ -th cluster is deleted, *i.e.*,  $\hat{Y}_r$  uses the weights  $w_{hik(gj)}a_{hik}$  instead of  $w_{hik(gj)}a_{hik(gj)}$ . The linearized version of  $v_J^*$  is the same as the variance estimator  $v_R$  (equation 3.4) with  ${}_c e_{hik}$  replaced by  $y_{hik}$  in the case of  $\hat{Y}_{ps}$ , and  $v_{JL}$  (equation 4.6) with  $e_{hik}$  replaced by  $y_{hik}$  in the case of the generalized regression estimator  $\hat{Y}_r$ . That is,

$$v_J^*(\hat{Y}_{ps}) = v(y_{hi}^*)$$

with

$$y_{hi}^* = \sum_c \sum_{k \in c_s} (n_h w_{hik}) y_{hik}$$

and

$$v_J^*(\hat{Y}_r) = v(y_{hi}^*)$$

with

$$y_{hi}^* = \sum_{k \in s} (n_h w_{hik}^*) y_{hik}.$$

Since  $v_J^*$  uses the  $y$ 's instead of the residuals  $e$ 's, it is clear that  $v_J^*$  should overestimate the true variance of the estimator, although it is computationally simpler than  $v_J$ .

### (i) Unconditional Results

To compare the unconditional performances of the variance estimators we computed the empirical relative bias (RB) for each variance estimator: RB of a variance estimator  $v$  is

$$RB = \frac{1}{MSE} \left[ \frac{1}{10,000} \sum_i v_i \right] - 1$$

where  $v_i$  is the value of  $v$  for the  $i$ -th simulated sample ( $i = 1, \dots, 10,000$ ) and MSE is the empirical MSE of the estimator, say  $\tilde{Y}$ :

$$MSE = \frac{1}{10,000} \sum_i (\tilde{Y}_i - Y)^2$$

where  $\tilde{Y}_i$  is the value of  $\tilde{Y}$  in the  $i$ -th simulated sample.

Error rates for normal theory confidence intervals on the total  $Y$  were also calculated for each variance estimator, using a nominal error rate of 5%:

error rate =

$$1 - \frac{1}{10,000} (\text{number of samples with } L_i \leq Y \leq U_i),$$

where  $L_i \leq Y \leq U_i$  is a confidence interval on  $Y$  for the  $i$ -th simulated sample. Lower and upper error rates were calculated as:

lower error rate =

$$\frac{1}{10,000} (\text{number of samples with } Y < L_i)$$

upper error rate =

$$\frac{1}{10,000} (\text{number of samples with } Y > U_i).$$

We also calculated the average lengths of the confidence intervals as

$$\text{average length} = \frac{1}{10,000} \sum_i (U_i - L_i).$$

Table 3 reports the unconditional results for the post-stratified estimator  $\hat{Y}_{ps}$  using the above performance measures. With respect to relative bias,  $v_{JL}$  and  $v_J$  both perform well with  $RB < 1\%$  while the incorrect jackknife  $v_{JL}^*$  severely overestimates the MSE ( $RB = 37\%$ ). We note that  $v_L$  is also estimating the MSE of  $\hat{Y}_{ps}$  well unconditionally ( $RB < 1\%$ ), contrary to Valliant's (1993) claim. Valliant (1993) reported RB of 35% for  $v_L$  using the same data set. In view of the design-consistency of  $v_L$  supplemented by our simulation results on  $v_L$ , we conjecture that Valliant's calculations on  $v_L$  might be incorrect.

**Table 3**

Unconditional Results for the Poststratified Estimator

Performance Measure	$v_L(\hat{Y}_{ps})$	$v_{JL}(\hat{Y}_{ps})$	$v_J(\hat{Y}_{ps})$	$v_J^*(\hat{Y}_{ps})$
Relative bias (%)	-0.44	0.12	0.26	37.16
Error rate (%)	5.20	5.09	5.06	2.41
Lower error rate (%)	2.41	2.35	2.33	0.99
Upper error rate (%)	2.79	2.74	2.73	1.42
Average length	3.81	3.82	3.83	4.48

Turning to confidence interval performance, Table 3 shows that the error rates associated with  $v_J$ ,  $v_{JL}$  and  $v_L$  are close to the nominal 5% while the error rate for  $v_J^*$  is considerably lower than 5% (about 2.5%). Performances with respect to lower and upper error rates are also similar. The variance estimators,  $v_J$ ,  $v_{JL}$  and  $v_L$ , perform similarly in terms of average length of confidence intervals while the average length associated with  $v_J^*$  is significantly larger due to overestimation bias. Finally, we note that the performance measures for  $v_J$  and  $v_{JL}$  are very close, supporting the asymptotic equivalence of  $v_J$  and  $v_{JL}$ .

**Table 4**

Unconditional Results for the Generalized Regression Estimator

Performance Measure	$v_L(\hat{Y}_r)$	$v_{JL}(\hat{Y}_r)$	$v_J(\hat{Y}_r)$	$v_J^*(\hat{Y}_r)$
Relative bias (%)	-0.96	0.76	0.57	25.87
Error rate (%)	5.30	5.27	5.23	3.07
Lower error rate (%)	2.24	2.21	2.19	1.08
Upper error rate (%)	3.06	3.06	3.04	1.99
Average length	3.94	3.95	3.95	4.44



Unconditional results for the generalized regression estimator  $\hat{Y}_r$  are reported in Table 4. As in the case of  $\hat{Y}_{ps}$ , the variance estimators  $v_J$ ,  $v_{JL}$  and  $v_L$  perform well both in terms of relative bias and error rates of confidence intervals. On the other hand, the incorrect jackknife  $v_J^*$  leads to severe overestimation which in turn is reflected in the lower than nominal error rates and larger average length of confidence intervals.

## (ii) Conditional Results

We have also studied conditional properties of the variance estimators, following Valliant (1993). For the poststratified estimator, we divided the 10,000 simulated samples into 10 groups each containing 1,000 samples using the measure (Valliant 1993)

$$D_{ps} = \sum_c \left( \frac{{}_c\hat{M}}{{}_cM} - 1 \right).$$

The measure  $D_{ps}$  was calculated for each sample and the 10,000 samples were sorted in ascending order according to the  $D_{ps}$ -values and then divided into groups. We may interpret  $D_{ps}$  as a measure of how “balanced” the sample is with respect to the distribution of the poststrata counts.

For the generalized regression estimator, we used the following natural extension of  $D_{ps}$ :

$$D_r = \sum_a \left( \frac{{}_a\hat{M}}{{}_aM} - 1 \right) + \sum_b \left( \frac{{}_b\hat{M}}{{}_bM} - 1 \right),$$

where  $a$  and  $b$  index the levels of the two poststratification variables and  $({}_a\hat{M}, {}_aM)$  and  $({}_b\hat{M}, {}_bM)$  are the corresponding marginal counts. We may interpret  $D_r$  as a measure of how “balanced” the sample is with respect to the distribution of the marginal poststrata counts.

**Table 5**

Conditional Relative Biases (%) for the Poststratified Estimator

Group	$v_L(\hat{Y}_{ps})$	$v_{JL}(\hat{Y}_{ps})$	$v_J(\hat{Y}_{ps})$	$v_J^*(\hat{Y}_{ps})$
1	-5.00	-8.05	-7.88	17.83
2	0.55	-1.18	-1.01	28.06
3	8.33	7.03	7.19	41.29
4	-1.10	-1.56	-1.42	31.82
5	-0.76	-0.69	-0.55	34.77
6	2.50	3.39	3.53	41.69
7	6.10	7.51	7.66	48.86
8	6.60	8.82	8.96	53.54
9	-4.46	-1.43	-1.31	41.11
10	-13.56	-9.17	-9.07	36.63

**Table 6**

Conditional Error Rates (%) for the Poststratified Estimator

Group	$v_L(\hat{Y}_{ps})$	$v_{JL}(\hat{Y}_{ps})$	$v_J(\hat{Y}_{ps})$	$v_J^*(\hat{Y}_{ps})$
1	5.5	5.9	5.9	3.4
2	4.6	4.8	4.8	2.9
3	3.7	3.8	3.8	1.9
4	5.7	5.8	5.8	2.9
5	4.9	4.8	4.7	2.6
6	5.1	5.0	4.8	2.2
7	5.2	4.8	4.8	2.1
8	4.5	4.3	4.3	1.3
9	5.8	5.4	5.4	2.4
10	7.0	6.3	6.3	2.4

The results for the poststratified estimator are given in Tables 5 and 6: conditional relative biases in Table 5 and conditional error rates (nominal 5%) in Table 6. These performance measures were computed in the same manner as the unconditional case but from each group separately. It is clear from Tables 5 and 6 that  $v_J$ ,  $v_{JL}$  and  $v_L$  all perform well, although  $v_L$  is somewhat worse in the extreme groups 1 and 10, while  $v_J^*$  performed poorly as before. It is somewhat surprising to see  $v_L$  performing so well conditionally. A possible explanation is that with our particular sampling design we have  $\hat{M} = \sum_{(hik) \in S} w_{hik} = M$  so that

$$\sum_c {}_c\hat{M} = \hat{M} = M.$$

Because of this, we do not obtain samples which are poorly balanced since if some poststrata counts  ${}_c\hat{M}$  are gross overestimates, say, then the other counts correct for the overestimation in order to satisfy the above constraint. Thus, we see mostly well balanced samples in which case  $v_L$  is expected to perform well.

**Table 7**

Conditional Relative Biases (%) for the Generalized Regression Estimator

Group	$v_L(\hat{Y}_r)$	$v_{JL}(\hat{Y}_r)$	$v_J(\hat{Y}_r)$	$v_J^*(\hat{Y}_r)$
1	9.25	4.95	5.13	26.51
2	3.99	1.50	1.67	24.96
3	-3.24	-4.76	-4.59	17.53
4	-2.66	-3.43	-3.26	20.53
5	7.90	7.61	7.80	35.46
6	-3.60	-3.12	-2.94	23.38
7	-9.24	-8.27	-8.08	17.41
8	3.34	5.30	5.50	35.84
9	-3.75	-0.85	-0.62	30.84
10	-8.68	-4.15	-3.92	28.50

**Table 8**

Conditional Error Rates (%) for the Generalized Regression Estimator

Group	$v_L(\hat{Y}_r)$	$v_{JL}(\hat{Y}_r)$	$v_J(\hat{Y}_r)$	$v_J^*(\hat{Y}_r)$
1	4.3	4.5	4.4	3.0
2	4.9	5.0	5.0	3.3
3	5.0	5.1	5.1	3.8
4	5.7	5.9	5.9	3.3
5	3.9	4.0	4.0	2.3
6	5.7	5.8	5.7	3.0
7	5.9	5.8	5.8	2.9
8	5.8	5.7	5.7	2.8
9	5.5	5.1	4.9	3.0
10	6.3	5.8	5.8	3.3

The results for the generalized regression estimator are given in Tables 7 and 8: conditional relative biases in Table 7 and conditional error rates (nominal 5%) in Table 8. The results are very similar to those for the one stratifier case. In both cases we again note that the performance measures for  $v_J$  and  $v_{JL}$  are very close, supporting the asymptotic equivalence of  $v_J$  and  $v_{JL}$ .

In summary, the three variance estimators  $v_J$ ,  $v_{JL}$  and  $v_L$  performed similarly. The incorrect jackknife  $v_J^*$  performed poorly indicating that reweighting must be done each time a cluster is deleted.

## 7. CONCLUDING REMARKS

Beebakhee (1995) applied the three variance estimators,  $v_J$ ,  $v_{JL}$  and  $v_L$ , to a number of household surveys conducted by Statistics Canada. Her empirical results showed that the jackknife linearization variance estimator,  $v_{JL}$ , consistently consumed less time and money for all study surveys than the jackknife variance estimator,  $v_J$ , and yet approximated  $v_J$  very well. These results are practically important because the users wanted a computationally simpler variance estimator which can approximate the currently used  $v_J$  very well. The standard linearization variance estimator  $v_L$  performed similar to  $v_{JL}$  in terms of cost and time, but it did not approximate  $v_J$  as well as  $v_{JL}$ .

If the primary interest is the estimation of totals or ratios, then the jackknife linearization variance estimator,  $v_{JL}$ , is attractive because it is computationally simpler than the jackknife variance estimator,  $v_J$ , and yet leads to values close to the jackknife. But for general smooth statistics  $v_{JL}$  suffers from the same disadvantage as the standard linearization variance estimator,  $v_L$ , in the sense that both require the derivation of a separate formula for each statistic, unlike  $v_J$ . In terms of statistical properties, our simulation study suggests that the three variance

estimators,  $v_J$ ,  $v_{JL}$ , and  $v_L$ , perform similarly. On the other hand, the incorrect jackknife  $v_J^*$ , which uses the same adjustment whenever a cluster is deleted, performs poorly indicating that reweighting must be done each time a cluster is deleted.

## ACKNOWLEDGEMENT

This work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

## APPENDIX

### Proof of the Result $v_J(\hat{Y}_r) \approx v_{JL}(\hat{Y}_r)$

To establish the desired result, we first approximate the difference  $\hat{A}_{(gj)}^{-1} - \hat{A}^{-1}$ . Using the matrix identity,

$$(I + PQ)^{-1} = I - P(I + QP)^{-1}Q$$

we get

$$\begin{aligned} \hat{A}_{(gj)}^{-1} - \hat{A}^{-1} &= \hat{A}^{-1}[I + (\hat{A}_{(gj)} - \hat{A})\hat{A}^{-1}]^{-1} - \hat{A}^{-1} \\ &= \hat{A}^{-1}[I - (\hat{A}_{(gj)} - \hat{A}) \\ &\quad (I + \hat{A}^{-1}(\hat{A}_{(gj)} - \hat{A}))^{-1}\hat{A}^{-1}] - \hat{A}^{-1} \\ &\approx -\hat{A}^{-1}(\hat{A}_{(gj)} - \hat{A})\hat{A}^{-1}. \end{aligned} \quad (A.1)$$

The approximation (A.1) follows by noting that (i)  $\hat{A}_{(gj)} - \hat{A}$  is of lower order than  $\hat{A}$  under the assumption that no cluster contribution is of disproportionate size as the number of strata  $L$  increases (see Yung (1996) for details on regularity conditions) and (ii)  $[I + \hat{A}^{-1}(\hat{A}_{(gj)} - \hat{A})]^{-1} \approx I - \hat{A}^{-1}(\hat{A}_{(gj)} - \hat{A})$ .

Using (A.1), we obtain

$$\begin{aligned} \hat{B}_{(gj)} - \hat{B} &= (\hat{A}_{(gj)}^{-1} - \hat{A}^{-1} + \hat{A}^{-1})(\hat{b}_{(gj)} - \hat{b} + \hat{b}) \\ &\quad - \hat{A}^{-1}\hat{b} \\ &\approx (\hat{A}_{(gj)}^{-1} - \hat{A}^{-1})\hat{b} + \hat{A}^{-1}(\hat{b}_{(gj)} - \hat{b}) \\ &\approx -\hat{A}^{-1}(\hat{A}_{(gj)} - \hat{A})\hat{B} + \hat{A}^{-1}(\hat{b}_{(gj)} - \hat{b}). \end{aligned} \quad (A.2)$$

It now follows from (A.2) that

$$\begin{aligned} \hat{Y}_{r(gj)} - \hat{Y}_r &\approx (\hat{Y}_{(gj)} - \hat{Y}) - (\hat{X}_{(gj)} - \hat{X})^T \hat{B} \\ &\quad - (\hat{X} - X)^T (\hat{B}_{(gj)} - \hat{B}) \\ &\approx \frac{1}{n_g - 1} (\bar{e}_g^* - e_{gj}^*), \end{aligned} \quad (A.3)$$



where  $e_{gj}^* = \sum_k (n_g w_{gjk}^*) e_{gjk}$  and  $\bar{e}_g^* = (1/n_g) \sum_j e_{gj}^*$ . We used the following results in arriving at (A.3):

$$(\hat{Y}_{(gj)} - \hat{Y}) - (\hat{X}_{(gj)} - \hat{X})^T \hat{B} = \frac{1}{n_g - 1} (\bar{e}_g - e_{gj})$$

and

$$(\hat{X} - X)^T (\hat{B}_{(gj)} - \hat{B}) \approx$$

$$(X - \hat{X})^T \hat{A}^{-1} \left[ \frac{1}{n_g - 1} (\bar{u}_g - u_{gj}) \right],$$

where  $e_{gj} = \sum_k (n_g w_{gjk}) e_{gjk}$  and  $u_{gj} = \sum_k (n_g w_{gjk}) x_{gjk} e_{gjk}$ .

It now follows from (A.3) that

$$\begin{aligned} v_J(\hat{Y}_r) &\approx \sum_{h=1}^L \frac{1}{n_h(n_h - 1)} \sum_{i=1}^{n_h} (e_{hi}^* - \bar{e}_h^*)^2 \\ &= v(e_{hi}^*) = v_{JL}(\hat{Y}_r). \end{aligned}$$

## REFERENCES

- BEEBAKHEE, R. (1995). A comparison of two variance estimation methods: The Jackknife and the linearized Jackknife. Methodology Branch Working Paper, HSMD-95-005E. Statistics Canada.
- BINDER, D.A. (1996). Linearization methods for single phase and two phase samples: A cookbook approach. *Survey Methodology*, 22, 17-22.
- CASADY, R.J., and VALLIANT, R. (1993). Conditional properties of poststratified estimators under normal theory. *Survey Methodology*, 19, 183-192.
- DEVILLE, J., and SÄRNDAL, C.E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- HUANG, E.T., and FULLER, W.A. (1978). Nonnegative regression estimation for sample survey data. *Proceedings of the Social Statistics Section, American Statistical Association*, 300-305.
- RAO, J.N.K. (1985). Conditional inference in survey sampling. *Survey Methodology*, 11, 15-31.
- ROYALL, R.M., and CUMBERLAND, W.G. (1981). An empirical study of the ratio estimator and estimator of its variance. *Journal of the American Statistical Association*, 76, 66-88.
- SÄRNDAL, C.E., SWENSSON, B., and WRETMAN, J. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76, 527-537.
- STATISTICS CANADA (1990). *Methodology of the Canadian Labour Force Survey*. Catalogue No. 71-526.
- VALLIANT, R. (1993). Poststratification and conditional variance estimation. *Journal of the American Statistical Association*, 88, 89-96.
- YUNG, W. (1996). Contributions to poststratification in stratified multi-stage samples. Unpublished Ph.D. thesis, Carleton University, Ottawa, Canada.





# Small Area Estimation Under an Inverse Gaussian Model

Y.P. CHAUBEY, F. NEBEBE and P.S. CHEN<sup>1</sup>

## ABSTRACT

In this paper, we consider analysis of variance methodology for inverse Gaussian distribution and adapt it for estimation of small area parameters in finite populations. It is demonstrated, through a Monte Carlo study, that these estimators offer a competitive choice for positively skewed survey data such as income or yield of a particular sector.

**KEY WORDS:** Interactions; Inverse Gaussian; Monte Carlo; Regression estimates; Synthetic estimates; Särndal-Hidiroglou estimator; Unbalanced model.

## 1. INTRODUCTION

Recently, a large number of methods appeared in the literature for the problem of small area estimation; for example Prasad and Rao (1990), Särndal and Hidiroglou (1989), Choudhry and Rao (1988), and Särndal (1984) and the references cited there, especially Särndal and Råbäck (1983), Fay and Herriot (1979), Schaible (1979), Holt, Smith and Tomberlin (1979), and Gonzalez and Hoza (1978), to name a few. The need for small area estimates of several characteristics of a given population has generated various useful procedures that produced realistic and sufficiently accurate estimates for local areas and other special subgroups. Several of the techniques suggested by the authors mentioned above were implicitly and/or explicitly model-based and utilized the standard normal theory. Others have tackled the provision of estimates for local areas from Bayesian and empirical Bayes perspectives by finding a compromise between the sample mean of an area (that is assumed to be normal) and an estimator based on regression on one or more covariates (see *e.g.*, Stroud 1987; MacGibbon and Tomberlin 1989). For an extensive review of recent developments in small area estimation, the reader may refer to Ghosh and Rao (1994).

The standard normal theory analysis of factorial experiments may be inappropriate to apply in situations where data are generated from markedly positively skewed distributions. While most of the inference procedures are analytically tractable, the accuracy and reliability of the results may be questionable in many practical applications. Thus, such an analysis based on positively skewed distributions is called for.

The objective of this paper is to consider inference procedures for unbalanced as well as balanced two-factor experiments under inverse Gaussian model that may be used to produce estimates for small regions. Hidiroglou and Särndal (1985) reported on a Monte Carlo study where a modified

regression estimator is preferred as a compromise between the synthetic estimator and the generalized regression estimator. Särndal and Hidiroglou (1989) also presented further comparisons of estimators on the basis of conditional inference. The generalized regression estimator is basically derived from a super population regression model without any distributional assumptions. Chaubey (1991) considered super population models of Durbin (1959) with gamma auxiliary and inverse Gaussian auxiliary in which case the generalized regression estimator has the property of being the best linear unbiased predictor (see Prasad and Rao 1990). In fact, the best linear unbiased predictor for the population total does not depend on the form of the distribution of the characteristic variable, hence this technique is preferable given that maximum likelihood estimates (MLE) may be hard to obtain. As we have seen that the super population distributions (as transfused in the populations) may resemble closely to inverse Gaussian distributions for variety of populations we would like to exploit this aspect of the population.

The use of inverse Gaussian distribution is not merely a superficial one but it has been used successfully in many situations (see Folks and Chhikara 1978) and resembles closely to gamma, log normal and Weibull populations which are common in modeling positively skewed non negative random variables. In this paper, we study the use of inverse Gaussian model in applying to the small area estimation. The approach of Fries and Bhattacharyya (1983) which discusses the analysis of two factor experiments under an inverse Gaussian model is of major importance. The above paper gives estimation in balanced, no-interaction model. We have extended this approach to unbalanced case, which is essential for estimation of domain totals or means. In this respect the general multiple regression approach of Bhattacharyya and Fries (1986), and Whitmore (1983) may be adapted, but we have chosen to take the direct approach. In Section 2 we specify the

<sup>1</sup> Y.P. Chaubey, Professor, Department of Mathematics and Statistics; F. Nebebe, Associate Professor, Department of Decision Sciences & M.I.S.; and P.S. Chen, Research Assistant, Department of Finance, Concordia University, Montreal, Canada.

model and present our proposed estimators under the inverse Gaussian model. In Section 3, a numerical study is carried out for evaluation of the performance of the proposed estimator through Monte Carlo simulation. Finally, Section 4 presents summary and conclusions.

## 2. THE INVERSE GAUSSIAN REGRESSION MODEL FOR SMALL AREA ESTIMATION

Suppose that a finite population  $\mathcal{U}$  is divided into  $D$  non-overlapping domains  $U_d$ ,  $d = 1(1)D$ , with  $N_d$  as the size of  $U_d$ . The population is further divided along a second dimension, into  $G$  non-overlapping groups  $U_g$ ,  $g = 1(1)G$ , with the size of  $U_g$  denoted by  $N_g$ . The cross-classification of domains and groups give rise to  $DG$  population cells  $U_{dg}$ ,  $d = 1(1)D$ ,  $g = 1(1)G$ , with  $N_{dg}$  as the size of  $U_{dg}$ . The population size  $N$  can then be expressed as  $N = \sum_d N_d = \sum_g N_g = \sum_{dg} N_{dg}$ . Our interest lies in estimating domain totals  $t_d = \sum_{U_d} y_k$ , where  $y$  represents the characteristic variable and  $y_k$  is the observation on  $k$ -th unit. A sample  $s$  of size  $n$  is selected from  $\mathcal{U}$  by a simple random sampling. Denote by  $s_d$ ,  $s_g$  and  $s_{dg}$  the parts of  $s$  that happen to fall in  $U_d$ ,  $U_g$  and  $U_{dg}$ . The corresponding sample sizes are denoted by  $n_d$ ,  $n_g$  and  $n_{dg}$ , respectively.

### 2.1 Regression Method for Inverse Gaussian Data

We refer readers to two recent comprehensive reviews about the developments in the inverse Gaussian distribution, namely, Chhikara and Folks (1989), and Iyengar and Patwardhan (1988). The probability density function of an inverse Gaussian variate with parameters  $(\theta, \sigma)$ ,  $IG(\theta, \sigma)$ , is given by

$$f(y; \theta, \sigma) = (2\pi\sigma)^{-1/2} y^{-3/2} \exp[-(2\sigma y)^{-1}(y\theta^{-1} - 1)^2]; \quad (2.1)$$

with  $y > 0$ ,  $\theta > 0$ ,  $\sigma > 0$ . The mean and variance of this distribution are  $\theta$  and  $\theta^3\sigma$ , respectively. Bhattacharyya and Fries (1982) proposed a reciprocal linear model for  $\theta$ . Specifically, they assume a model of the form  $\theta_k^{-1} = x_k'\eta$ . An estimator of  $\eta$ , similar to the estimator of the regression parameter in the usual linear model (see Särndal 1984) in this situation is given by

$$\hat{\eta} = \left( \sum_{k \in S_d} \frac{x_k x_k' y_k}{\pi_k} \right)^{-1} \sum_{k \in S_d} \frac{x_k}{\pi_k}. \quad (2.2)$$

This is called pseudo Maximum Likelihood estimator, because it is obtained by unconditional maximization of the likelihood function and therefore  $x_k'\hat{\eta} > 0$  may not be satisfied for all  $k$ . Then an estimator of the total  $t_d$  of

the  $d$ -th domain in the spirit of Särndal's (1984) modified regression estimator may be constructed as

$$\hat{t}_{dIG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in S_d} \frac{e_k}{\pi_k} \quad (2.3)$$

where  $\hat{y}_k = x_k'\hat{\eta}$  and  $e_k = y_k - \hat{y}_k$ . In what follows, we denote the mean of the  $(d, g)$  cell by  $\theta_{dg}$ , and consider the case of simple random sampling in which case  $\pi_k$ 's are constant. We first discuss the prediction of observations for the use of (2.3) based on an additive effects model given by,

$$\theta_{dg}^{-1} = \mu + \alpha_d + \beta_g, \quad \sum \alpha_d = \sum \beta_g = 0, \quad (2.4)$$

where  $\mu$ ,  $\alpha_d$ 's and  $\beta_g$ 's represent the overall effect, the domain or row effects, and the group or column effects, respectively. For the inverse Gaussian distribution we must also have  $\theta_{dg} > 0$  for all  $(d, g)$  and  $\sigma > 0$ . Thus the parameters  $\mu$ ,  $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_D)$ ,  $\beta = (\beta_1, \beta_2, \dots, \beta_G)$ , and  $\sigma$  lie in the set  $\Omega = \{(\mu, \alpha, \beta, \sigma) : \sum_d \alpha_d = 0, \sum_g \beta_g = 0; \mu + \alpha_d + \beta_g > 0, \forall (d, g); \sigma > 0\}$ . Under this setup estimation of parameters for prediction can be accomplished through unconditional maximization of the likelihood function. Conditional on the population and the sample sizes  $n_{dg}$  and referring to (2.1) and (2.3), the log-likelihood function of the parameters is given by

$$\ell = -\frac{1}{2} \log \sigma \sum_d \sum_g n_{dg} - (2\sigma)^{-1} \sum_d \sum_g \sum_k y_{dgk}^{-1} [y_{dgk}(\mu + \alpha_d + \beta_g) - 1]^2. \quad (2.5)$$

We first note that the parameters are effectively given by  $(\mu, \alpha_d, \beta_g, d = 1, 2, \dots, D-1; g = 1, 2, \dots, G-1)$ . Thus, differentiating the above with respect to  $(\mu, \alpha_d, \beta_g, d = 1, 2, \dots, D-1; g = 1, 2, \dots, G-1)$  and equating the resulting partial derivatives to zero gives the following equations for the estimators  $(\hat{\mu}, \hat{\alpha}_d, \hat{\beta}_g, d = 1, 2, \dots, D-1; g = 1, 2, \dots, G-1)$ ,

$$\begin{aligned} \hat{\mu} y_{..} + \sum_{d=1}^{D-1} \hat{\alpha}_d (y_{d.} - y_{D.}) + \sum_{g=1}^{G-1} \hat{\beta}_g (y_{.g} - y_{.G}) &= n_{..}, \\ \hat{\mu} (y_{d.} - y_{D.}) + \hat{\alpha}_d y_{d.} + \sum_{j=1}^{D-1} \alpha_j y_{Dj} \\ &+ \sum_{g=1}^{G-1} \hat{\beta}_g \{ (y_{dg} - y_{Dg}) - (y_{dG} - y_{DG}) \} = n_{d.} - n_{D.}, \\ \hat{\mu} (y_{.g} - y_{.G}) + \sum_{d=1}^{D-1} \hat{\alpha}_d \{ (y_{dg} - y_{dG}) - (y_{DG} - y_{DG}) \} \\ &+ \hat{\beta}_g y_{.g} + \sum_{j=1}^{G-1} \hat{\beta}_j y_{.G} = n_{.g} - n_{.G}, \end{aligned} \quad (2.6)$$



where the totals and means are represented by the notations

$$y_{dg} = \sum_k y_{dgk}, y_d = \sum_g y_{dg}, y_{.g} = \sum_d y_{dg}, \quad (2.7a)$$

$$n_d = \sum_g n_{dg}, n_{.g} = \sum_d n_{dg}, n_{..} = \sum_d \sum_g n_{dg}. \quad (2.7b)$$

The solutions  $(\hat{\mu}, \hat{\alpha}_d, \hat{\beta}_g)$ ,  $d = 1(1)D$ ,  $g = 1(1)G$ , provide the pseudo Maximum Likelihood estimator and may not yield nonnegative response estimates but will coincide with proper MLE as  $n_{dg} \rightarrow \infty$  (see Fries and Bhattacharyya 1983) with probability one. Negative values of the response estimates may thus be truncated to zero.

In the case of the  $IG(\theta, \sigma)$  model with interaction, the usual parameterization of the interaction effects suggests the model

$$\theta_{dg}^{-1} = \mu + \alpha_d + \beta_g + \gamma_{dg},$$

$$\sum_d \alpha_d = \sum_g \beta_g = \sum_d \gamma_{dg} = \sum_g \gamma_{dg} = 0, \quad (2.8)$$

where now  $\gamma_{dg}$  is the interaction effect when domain is at the  $d$ -th level and group is at the  $g$ -th level. The estimators of parameters may be obtained in this case following the method outlined above. However, noting that the maximum likelihood estimator (MLE) of  $\theta_{dg}$  is  $\bar{y}_{dg}$  and there is one to one relation between the parameters in the reparametrized model in terms of  $(\mu, \alpha_d, \beta_g, \gamma_{dg})$  and the original parameters  $\theta_{dg}$ , explicit formulae for the MLE of different parameters are not needed. Corresponding to equation (2.3), therefore, for a two-factor model with interaction, our estimator is

$$\hat{t}_{dWI} = \sum_g N_{dg} \bar{y}_{dg}, \quad (2.9)$$

which is the post stratified estimator and is not of further interest in small area estimation. For the model without interaction, the estimator is given as

$$\hat{t}_{dWOI} = \sum_g N_{dg} \hat{\theta}_{dg} + \sum_g \hat{N}_{dg} (\bar{y}_{dg} - \hat{\theta}_{dg}), \quad (2.10)$$

where  $\hat{\theta}_{dg}^{-1} = \hat{\mu} + \hat{\alpha}_d + \hat{\beta}_g$ , the estimators being obtained from (2.6) and  $\hat{N}_{dg} = n_{dg}N/n_{..}$ .

In order to judge the effectiveness of this estimator a numerical study has been performed and is reported in the following section.

### 3. A NUMERICAL STUDY OF THE INVERSE GAUSSIAN REGRESSION ESTIMATOR

In this section we provide the results of a simulation study which evaluates the performance of the estimators developed in the previous section. The modified regression estimator due to Särndal and Hidiroglou (1989) given below will be used as the bench mark for the above purpose;

$$\hat{t}_{ds-H} = \sum_g N_{dg} \bar{y}_{.g} + \sum_g F_d \hat{N}_{dg} (\bar{y}_{dg} - \bar{y}_{.g}), \quad (3.1)$$

where  $F_d = N_d/\hat{N}_d$  if  $\hat{N}_d \geq N_d$ , otherwise  $F_d = \hat{N}_d/N_d$ . Here,  $\hat{N}_d = n_d N/n_{..}$ . An alternative form of this estimator which takes into account both group and domain effects can be obtained by replacing  $\bar{y}_{.g}$  by  $\bar{y}_{.g} + \bar{y}_d - \bar{y}_{..}$  but this has not been pursued here. It should be noted that the above estimators cannot be computed when  $n_{dg}$  is zero. When this happens the estimators are simply taken to be the sample means of the respective domains. We also include the following modified version of  $\hat{t}_{dWOI}$ ,

$$\hat{t}_{dWOIM} = \sum_g N_{dg} \hat{\theta}_{dg} + \sum_g F_d \hat{N}_{dg} (\bar{y}_{dg} - \hat{\theta}_{dg}), \quad (3.2)$$

for comparison.

#### 3.1 Design of the Simulation Study

We consider Household Income data for Canadians in 1986, obtained from Household Income, Facilities and Equipment microdata tape of Statistics Canada (1987), for generating the values of parameters to be used for simulation. Using Household incomes, from these data, dividing them into 10 provinces and 6 educational groups, we first fit an inverse Gaussian model given by equation (2.4). The estimates of parameters are then used in forming the true parameters of the inverse Gaussian super population model which are summarized in appendix A. The values of  $D$ ,  $G$ ,  $N_{dg}$  are chosen from this population (see appendix B), where  $D$  represents the number of provinces (*i.e.*,  $D = 10$ ) and  $G$  represents the number of education groups (*i.e.*,  $G = 6$ ). Further sets of values of  $\theta_{dg}$  and  $\sigma$  are obtained by considering various combinations of  $(c_1, c_2)$ ;  $c_1 = 0(1)4$  and  $c_2 = 1, .25, .1, .01$  where  $c_1$  is used to transform  $\theta_{dg}$  to  $10^{-c_1}\theta_{dg}$  and  $c_2$  is used to transform  $\sigma$  to  $c_2\sigma$ . Note that  $c_1 = 0$  and  $c_2 = 1$  gives the parameter values for the original population. Also, the higher values of  $c_1$  indicate smaller values of the means and those of  $c_2$  indicate higher value of the dispersion parameter.

For the simulation study, first we generate for a given set of  $\theta_{dg}$  and  $\sigma$  values an inverse Gaussian random sample using the algorithm in Michael *et al.* (1976) with number of observations according to the values given in

the appendix B. This random sample is then used as a finite population from which we select 1000 random samples for each of the sample fractions, 1%, and 5% with replacement. We had actually selected several random samples and obtained similar results as reported here. From each sample we computed the estimators of totals for the 10 domains using estimators  $\hat{t}_{dS-H}$ ,  $\hat{t}_{dWOI}$  and  $\hat{t}_{dWOIM}$ . The criteria for evaluating the performance of the estimators are the mean absolute relative error (MARE) and the absolute relative bias (ARB) defined as follows:

$$\text{MARE}(\hat{t}_d) = \frac{1}{1000} \sum_{i=1}^{1000} |\hat{t}_{di} - t_d| / t_d \quad (3.3)$$

$$\text{ARB}(\hat{t}_d) = \left| \frac{1}{1000} \sum_{i=1}^{1000} \hat{t}_{di} - t_d \right| / t_d. \quad (3.4)$$

Here  $\hat{t}_d$  denotes a typical estimator of  $t_d$  and  $\hat{t}_{di}$  denotes the value of the  $i$ -th Monte Carlo sample ( $i = 1, \dots, 1000$ ).

**Table 1**  
Mean Absolute Relative Error (%) of Different Estimators

Domain	1% Sample			5% Sample			1% Sample			5% Sample		
	SH	WOI	WOIM	SH	WOI	WOIM	SH	WOI	WOIM	SH	WOI	WOIM
$c_1 = 0, c_2 = 1$							$c_1 = 0, c_2 = .01$					
1	13.27	13.05	13.19	6.60	6.48	6.47	3.72	2.46	2.45	1.80	0.89	0.89
2	14.57	13.61	14.20	7.53	7.61	7.69	3.79	3.56	3.48	2.10	0.59	0.60
3	25.27	27.86	26.88	19.07	20.74	20.80	2.52	1.51	1.52	1.19	0.77	0.77
4	11.83	11.70	11.74	5.29	5.61	5.59	1.83	1.08	1.09	0.93	0.58	0.58
5	10.57	11.72	11.68	6.80	7.10	7.11	0.92	0.90	0.91	0.42	0.40	0.40
6	7.12	7.45	7.52	3.85	3.95	3.97	1.94	1.22	1.22	0.93	0.64	0.64
7	11.78	13.91	14.23	7.39	8.01	8.05	1.22	1.13	1.14	0.86	0.64	0.64
8	11.48	12.56	12.46	6.70	7.15	7.14	1.29	0.93	0.94	0.76	0.67	0.68
9	7.43	7.92	7.99	3.61	3.74	3.75	3.47	2.99	2.96	3.13	2.97	2.96
10	15.32	17.43	17.16	11.20	11.81	11.80	0.93	0.94	0.95	0.52	0.52	0.53
$c_1 = 2, c_2 = 1$							$c_1 = 2, c_2 = .01$					
1	3.34	2.18	2.15	1.66	0.79	0.78	2.99	1.48	1.44	1.47	0.08	0.08
2	4.14	3.94	3.82	2.14	1.07	1.06	0.54	3.37	3.27	1.86	0.14	0.13
3	2.44	1.67	1.65	1.17	0.71	0.70	1.81	0.45	0.44	0.87	0.07	0.07
4	2.05	1.70	1.69	0.98	0.70	0.70	1.32	0.36	0.35	0.66	0.07	0.07
5	1.08	1.17	1.16	0.50	0.51	0.51	0.27	0.13	0.13	0.11	0.05	0.05
6	1.74	1.14	1.14	0.78	0.52	0.52	1.29	0.13	0.13	0.55	0.05	0.05
7	1.90	1.57	1.56	0.91	0.72	0.72	1.22	0.31	0.31	0.56	0.07	0.07
8	1.48	1.38	1.38	0.70	0.60	0.60	0.81	0.18	0.18	0.38	0.06	0.06
9	1.41	1.30	1.29	0.67	0.59	0.58	0.69	0.14	0.14	0.30	0.06	0.06
10	1.22	1.38	1.38	0.56	0.59	0.59	0.26	0.15	0.15	0.10	0.06	0.06
$c_1 = 4, c_2 = 1$							$c_1 = 4, c_2 = .01$					
1	2.99	1.48	1.44	1.47	0.08	0.08	2.99	1.45	1.41	1.47	0.01	0.01
2	3.54	3.37	3.27	1.86	0.14	0.13	3.54	3.36	3.25	1.87	0.05	0.05
3	1.81	0.45	0.44	0.87	0.07	0.07	1.80	0.38	0.37	0.86	0.01	0.01
4	1.32	0.36	0.35	0.66	0.07	0.07	1.31	0.28	0.27	0.66	0.01	0.01
5	0.27	0.13	0.13	0.11	0.05	0.05	0.24	0.06	0.06	0.10	0.01	0.01
6	1.29	0.13	0.13	0.55	0.05	0.05	1.29	0.06	0.06	0.54	0.01	0.01
7	1.22	0.31	0.31	0.56	0.07	0.07	1.20	0.24	0.24	0.55	0.01	0.01
8	0.81	0.18	0.18	0.38	0.06	0.06	0.79	0.09	0.09	0.37	0.01	0.01
9	0.69	0.14	0.14	0.30	0.06	0.06	0.68	0.06	0.06	0.29	0.01	0.01
10	0.26	0.15	0.15	0.10	0.06	0.06	0.23	0.07	0.07	0.09	0.01	0.01



**Table 2**  
Absolute Relative Bias (%) of Different Estimators

Domain	1% Sample			5% Sample			1% Sample			5% Sample		
	SH	WOI	WOIM	SH	WOI	WOIM	SH	WOI	WOIM	SH	WOI	WOIM
$c_1 = 0, c_2 = 1$							$c_1 = 0, c_2 = .01$					
1	4.34	2.40	2.51	1.87	0.26	0.27	2.66	1.58	1.54	1.22	0.03	0.03
2	8.88	3.46	4.39	2.18	0.30	0.23	3.15	3.40	3.31	1.38	0.04	0.04
3	3.13	3.47	2.74	0.51	1.12	1.15	1.44	0.31	0.32	0.68	0.01	0.01
4	1.57	0.51	0.53	0.50	0.21	0.22	1.11	0.29	0.30	0.53	0.03	0.03
5	0.13	0.33	0.35	0.20	0.16	0.18	0.10	0.03	0.02	0.05	0.01	0.01
6	1.09	0.14	0.04	0.02	0.39	0.42	1.09	0.03	0.03	0.43	0.02	0.01
7	1.20	1.09	1.59	0.54	0.28	0.30	0.99	0.22	0.23	0.43	0.01	0.01
8	0.40	0.04	0.12	0.20	0.53	0.54	0.55	0.00	0.01	0.28	0.03	0.03
9	1.03	0.47	0.36	0.24	0.04	0.01	1.01	0.35	0.37	0.45	0.14	0.14
10	1.05	2.27	2.03	0.04	0.30	0.29	0.08	0.02	0.01	0.06	0.01	0.01
$c_1 = 2, c_2 = 1$							$c_1 = 2, c_2 = .01$					
1	2.40	1.37	1.33	1.13	0.01	0.01	2.47	1.43	1.39	1.15	0.01	0.01
2	3.00	3.28	3.16	1.33	0.02	0.01	3.06	3.34	3.24	1.36	0.03	0.03
3	1.53	0.39	0.38	0.70	0.04	0.04	1.46	0.35	0.34	0.65	0.01	0.01
4	1.00	0.25	0.25	0.53	0.04	0.04	1.01	0.23	0.23	0.49	0.00	0.00
5	0.10	0.02	0.03	0.04	0.00	0.01	0.10	0.01	0.02	0.04	0.00	0.00
6	1.16	0.01	0.01	0.47	0.02	0.02	1.15	0.01	0.00	0.46	0.00	0.00
7	1.00	0.27	0.27	0.42	0.00	0.00	0.95	0.21	0.21	0.41	0.00	0.00
8	0.48	0.04	0.04	0.25	0.01	0.01	0.57	0.04	0.04	0.26	0.00	0.00
9	0.64	0.06	0.05	0.27	0.02	0.02	0.61	0.01	0.00	0.26	0.00	0.00
10	0.01	0.02	0.02	0.02	0.00	0.00	0.06	0.01	0.01	0.03	0.00	0.00
$c_1 = 4, c_2 = 1$							$c_1 = 4, c_2 = .01$					
1	2.47	1.43	1.39	1.15	0.01	0.01	2.48	1.43	1.39	1.15	0.00	0.00
2	3.06	3.34	3.24	1.36	0.03	0.03	3.07	3.35	3.24	1.36	0.04	0.04
3	1.46	0.35	0.34	0.65	0.01	0.01	1.45	0.34	0.34	0.64	0.00	0.00
4	1.01	0.23	0.23	0.49	0.00	0.00	1.01	0.24	0.24	0.49	0.00	0.00
5	0.10	0.01	0.02	0.04	0.00	0.00	0.11	0.01	0.02	0.04	0.00	0.00
6	1.15	0.01	0.00	0.46	0.00	0.00	1.15	0.01	0.00	0.46	0.00	0.00
7	0.95	0.21	0.21	0.41	0.00	0.00	0.94	0.20	0.20	0.41	0.00	0.00
8	0.57	0.04	0.04	0.26	0.00	0.00	0.58	0.04	0.05	0.26	0.00	0.00
9	0.61	0.01	0.00	0.26	0.00	0.00	0.60	0.00	0.00	0.25	0.00	0.00
10	0.06	0.01	0.01	0.03	0.00	0.00	0.06	0.01	0.01	0.03	0.00	0.00

### 3.2 Analysis of Results

The MARE values computed according to (3.3) and the ARB values from (3.4) for the three estimators and for different sample sizes are reported in Tables 1 and 2, respectively for a selection of pairs  $(c_1, c_2)$ . The values of  $c_1$  are chosen to represent, large means (as in the original population,  $c_1 = 0$ ), moderate means ( $c_1 = 2$ ) and small means ( $c_1 = 4$ ), whereas, the values chosen for  $c_2$  represent the original dispersion parameter ( $c_2 = 1$ ) and a further smaller value ( $c_2 = .01$ ). It may

be interesting to note that increasing  $c_1$  by 1 while keeping  $c_2$  fixed reduces the coefficient of variation by a factor of 10.

Some of the MARE and ARB values reported in Tables 1 and 2 are also plotted for visual inspection in Figures 1 and 2 for 1% samples, respectively.

When comparing the MARE and ARB values, reductions in biases as well as in relative errors are observed in many cases for both 1% and 5% samples. It is found that, the MARE and ARB values decrease with decreasing values of mean and dispersion parameter  $\sigma$ . Reductions

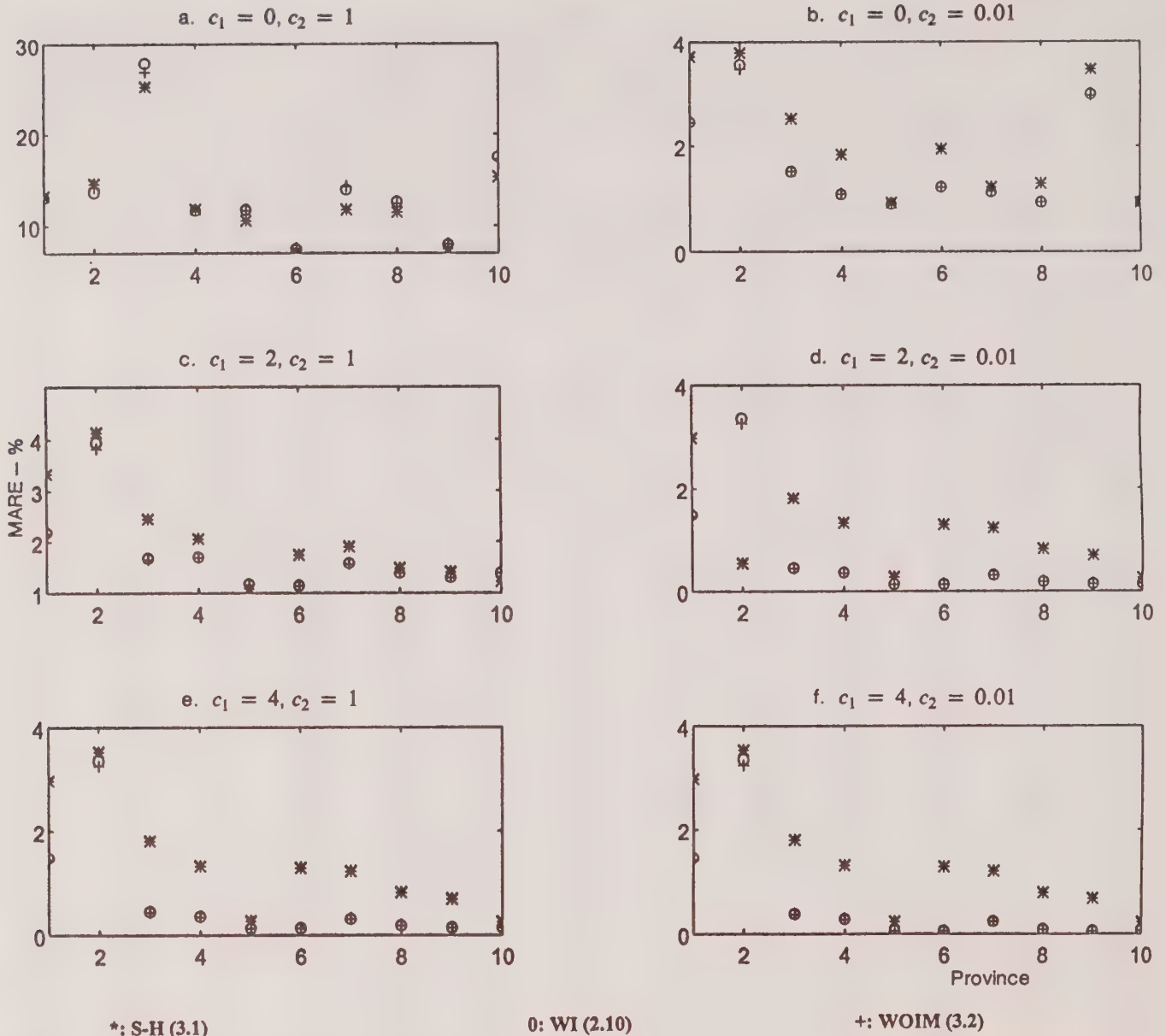


Figure 1. Mean absolute relative errors for different estimators for 1% sample.

are substantial, especially in case of 5% sample and/or when means are small. Note also that the reductions in bias are generally larger than reductions in the errors. We may note from Johnson and Kotz (1970, p. 141) that for fixed value of the mean, the standardized inverse Gaussian distribution tends to unit normal as the coefficient of variation tends to zero. Since larger gains in MARE and ARB values are noted for small values of the coefficient of variation, we conclude that proper modeling of the mean is important when the coefficient of variation is small for model based estimation.

We further find that  $\hat{t}_{dWOI}$  and  $\hat{t}_{dWOIM}$  have almost same MARE and ARB which indicates that the modification

of the estimator in (2.10) is not necessary. It may be remarked that the estimator  $\hat{t}_{dS-H}$ , in contrast, has been demonstrated (see Hidiroglou and Särndal 1985) to be substantial improvement over the corresponding unmodified estimator due to Särndal (1984).

Owing to the criticism of  $\hat{t}_{dWOI}$  and  $\hat{t}_{dWOIM}$  as being model dependent, we want to defend these on the following grounds. The inverse Gaussian distribution offers a variety of shapes and may be able to approximate lognormal, gamma, Weibull and such other positively skewed shapes. If we suspect that the principal characteristic is positively skewed, then the methodology we discussed here is viable and useful.



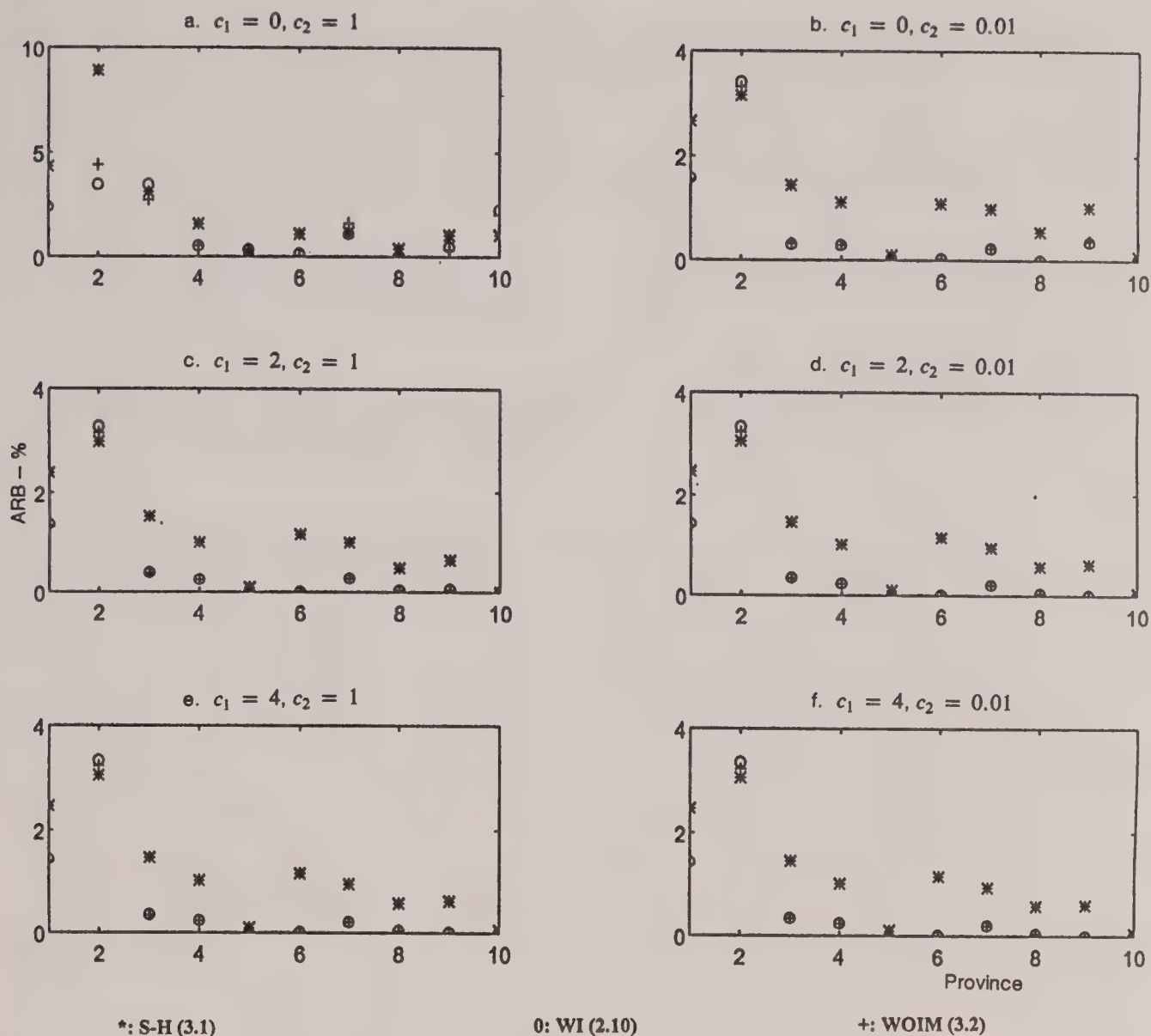


Figure 2. Absolute relative biases for different estimators for 1% sample.

#### 4. SUMMARY AND CONCLUSIONS

The generalization of analysis of variance methodology for inverse Gaussian population for unbalanced design was considered. The models without interactions of factors were studied and applied to the problem of estimation of small area parameters in finite populations. Using Canadian survey data, synthetic populations were generated in a Monte Carlo study. Through this we demonstrated that the proposed estimators perform well under a variety of conditions when the population can be regarded as a random sample from some inverse

Gaussian distribution. This approach offers a competitive choice for estimation of parameters in positively skewed survey data.

#### ACKNOWLEDGEMENTS

Support to Y.P. Chaubey and F. Nebebe by the Natural Sciences and Engineering Research Council of Canada is gratefully acknowledged. The authors thank the Editor, Professors J.N.K. Rao and A.B. Sim and the referees for their helpful comments.

**APPENDIX A****Values of the Parameters for Generation of the IG Population**

$$\mu = 3.13241147 \times 10^{-5}, \quad \sigma = 2.5447984 \times 10^{-5}$$

$d$	1	2	3	4	5
$10^6 \times \alpha_d$	3.1902855	2.8235779	1.5676078	.8056079	-.95350458

$d$	6	7	8	9	10
$10^6 \times \alpha_d$	-4.0661125	.49944356	.0061694263	-2.7414128	-1.1316622

$g$	1	2	3	4	5	6
$10^5 \times \beta_g$	1.0938451	.36781639	-.012707035	-.11561414	-.30936835	-1.023972

$\theta_{dg}$  values:

$d/g$	1	2	3	4	5	6
1	22,000.82	26,183.11	29,080.48	29,977.59	31,826.13	41,195.19
2	22,179.76	26,436.94	29,393.94	30,310.79	32,201.96	41,827.05
3	22,815.33	27,344.90	30,520.70	31,510.37	33,559.25	44,146.20
4	23,219.00	27,926.81	31,247.41	32,285.58	34,439.96	45,682.95
5	24,207.76	29,369.63	33,064.91	34,229.61	36,661.02	49,674.90
6	26,180.44	32,324.63	36,858.30	38,311.45	41,383.33	58,760.34
7	23,385.24	28,167.65	31,549.24	32,607.90	34,806.97	46,330.96
8	23,658.15	28,564.53	32,047.98	33,140.96	35,415.03	47,414.57
9	25,302.90	30,997.31	35,142.43	36,461.01	39,232.58	54,516.76
10	24,312.62	29,524.12	33,260.85	34,439.64	36,902.04	50,118.45

**APPENDIX B****Values of the Cell Sizes  $N_{dg}$** 

$d/g$	1	2	3	4	5	6	Total
1	627	360	277	84	215	110	1,673
2	285	212	198	72	68	83	918
3	597	483	616	148	204	231	2,279
4	729	397	568	151	239	219	2,303
5	1,372	761	1,216	202	473	511	4,535
6	1,177	888	1,795	517	707	800	5,884
7	639	432	673	165	236	222	2,367
8	850	512	888	264	349	297	3,160
9	700	699	1,350	385	696	572	4,401
10	456	540	1,083	342	393	407	3,221



## REFERENCES

- BHATTACHARYYA, G.K., and FRIES, A. (1986). On the inverse Gaussian multiple regression and model checking procedures. In *Reliability and Quality Control*, (A.P. Basu, Ed.). New York: North Holland, 86-100.
- CHAUBEY, Y.P. (1991). A study of ratio and product estimators under super population. *Communications in Statistics*, A, 20 (5 and 6), 1731-1746.
- CHOUDHRY, G.H., and RAO, J.N.K. (1988). Evaluation of small area estimators: An empirical study. Paper presented at the International Symposium on Small Area Statistics, New Orleans.
- CHHIKARA, R.S., and FOLKS, J.L. (1989). *The Inverse Gaussian Distribution*. New York: Marcel Dekker, Inc.
- DURBIN, J. (1959). A Note on the application of Quenouille's method of bias reduction to the estimation of ratio. *Biometrika*, 46, 477-480.
- FAY, R.E., and HERRIOT, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.
- FOLKS, J.L., and CHHIKARA, R.S. (1978). The inverse Gaussian distribution and its statistical applications-A review. *Journal of the Royal Statistical Society, Series B*, 40, 263-275.
- FRIES, A., and BHATTACHARYYA, G.K. (1983). Analysis of two-factor experiments under an inverse Gaussian model. *Journal of the American Statistical Association*, 78, 820-826.
- GONZALES, M.E., and HOZA, C. (1978). Small area estimation with application to unemployment and housing estimates. *Journal of the American Statistical Association*, 73, 7-15.
- GHOSH, M., and RAO, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 55-76.
- HIDIROGLOU, M.A., and SÄRNDAL, C.-E. (1985). An empirical study of some regression estimations for small domains. *Survey Methodology*, 6, 65-77.
- HOLT, D., SMITH, T.M.F., and TOMBERLIN, T.J. (1979). A model-based approach to estimation for small subgroups of a population. *Journal of the American Statistical Association*, 74, 405-410.
- IYENGAR, S., and PATWARDHAN, G. (1988). Recent developments in the inverse Gaussian distribution. In *Handbook of Statistics*, New York: Elsevier Science 479-490.
- JOHNSON, N. L., and KOTZ, S. (1970). *Continuous Univariate Distributions-I, Distributions in Statistics*. New York: Wiley.
- MACGIBBON, B., and TOMBERLIN, T.J. (1989). Small area estimates of proportions via empirical Bayes techniques. *Survey Methodology*, 15, 237-252.
- MICHAEL, J.R., SCHUCANY, W.R., and HASS, R.W. (1976). Generating random variables using transformations with multiple roots. *American Statistician*, 30(2), 88-90.
- PRASAD, N.G.N., and RAO, J.N.K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, 85, 163-171.
- PURCELL, N.J., and KISH, L. (1980). Postcensal estimates for local areas (or domains). *Bulletin of the International Statistical Institute*, 48, 3-18.
- SÄRNDAL, C.-E. (1984). Design consistent versus model dependent estimation for small domains. *Journal of the American Statistical Association*, 79, 624-631.
- SÄRNDAL, C.-E., and HIDIROGLOU, M.A. (1989). Small domain estimation: A conditional analysis. *Journal of the American Statistical Association*, 84, 266-275.
- SÄRNDAL, C.-E., and RÅBÄCK, G. (1983). Variance reduction and unbiasedness for small domain estimators. *Statistical Review*, 21, (Essays in Honor of T.E. Dalenius), 33-40.
- SCHAIBLE, W.L. (1979). A composite estimator for small area statistics. In *Synthetic Estimates for Small Areas*, NIDA Research Monograph 24, (J. Steinberg, Ed.). Rockville, MD: National Institute on Drug Abuse, 36-53.
- STATISTICS CANADA (1987). Microdata file, Household Income, Facilities and Equipment (1987), Statistics Canada, Household Surveys Division.
- STROUD, T.W.F. (1987). Bayes and empirical Bayes approaches to small area estimation. In *Small Area Statistics*, (R. Platek, J.N.K. Rao, C.-E. Särndal and M.P. Singh, Eds.). New York: Wiley, 124-137.
- WHITMORE, G.A. (1983). A regression method for censored inverse Gaussian data. *The Canadian Journal of Statistics*, 11, 305-315.





# A Comparison of Some Weighting Adjustment Methods for Panel Nonresponse

LOU RIZZO, GRAHAM KALTON and J. MICHAEL BRICK<sup>1</sup>

## ABSTRACT

In some surveys, many auxiliary variables are available for respondents and nonrespondents for use in nonresponse adjustment. One decision that arises is how to select which of the auxiliary variables should be used for this purpose and another decision involves how the selected variables should be used. Several approaches to forming weighting adjustments for nonresponse are considered in this research. The methods include those based on logistic regression models, categorical search algorithms, and generalized raking. These methods are applied to adjust for panel nonresponse in the Survey of Income and Program Participation (SIPP). The estimates from the alternative adjustments are assessed by comparing them to one another and to benchmark estimates from other sources.

**KEY WORDS:** Nonresponse bias; Panel surveys; Generalized raking; Benchmark estimates.

## 1. INTRODUCTION

Weights are commonly used in the analysis of survey data to compensate for unequal selection probabilities of the sampled elements, to compensate for unit nonresponse, and to make the weighted sample distributions for certain variables conform to known population distributions for those variables (thereby aiming to compensate for non-coverage and to improve the precision of the survey estimates) (Kish 1992). Corresponding to these three objectives, the weights are usually developed in three stages. First, a base weight is calculated for each sampled element as the inverse of the element's selection probability. Second, the base weights of responding sampling elements are multiplied by a nonresponse weight to compensate for the nonrespondents. Third, the adjusted weight is modified to make the weighted sample distributions for certain variables conform to external information on these distributions.

This paper deals with the nonresponse adjustment weights that attempt to compensate for unit nonresponse. A commonly used procedure for obtaining these weights is to divide the total sample into a set of weighting classes based on information known for both respondents and nonrespondents, and then to increase the base weights for the respondents in a weighting class to represent the nonrespondents in that class (Oh and Scheuren 1983; Kalton 1983). In many surveys little information is known about the nonrespondents, beyond the primary sampling units and strata from which they come. In this case, the choice of possible weighting classes is limited, and the procedure can be applied fairly straightforwardly.

In some surveys, however, there is an extensive amount of information available for the nonrespondents. This information may be available from the sampling frame

(*e.g.*, when sampling employees from personnel files) or by matching sampled elements with administrative records. Also, in panel surveys and other surveys involving more than one stage of data collection, extensive information on nonrespondents at later stages is available from their responses at the early stages.

The major focus of this research is on methods for developing weighting adjustments for nonresponse when a large number of characteristics of the nonrespondents are known. In this situation, decisions about methods of adjusting for nonresponse involve selecting which auxiliary variables will be used and how they will be used to make the adjustments.

The main ideas are presented in this article by applying several different adjustment procedures in a specific panel survey, the Survey of Income and Program Participation (SIPP). The SIPP is an ongoing household panel survey conducted by the U.S. Bureau of the Census. The nonrespondents to a SIPP panel can be separated in two groups: those who fail to respond at the initial wave of data collection (initial wave nonrespondents), and those who respond at the initial wave but fail to respond at one or more of the subsequent waves of the panel for which they are eligible (panel nonrespondents). For the latter group, extensive information from the initial wave of data collection can be utilized in adjusting for panel nonresponse. The weighting adjustments studied here relate to the panel nonrespondents only. These adjustments modify the weights of panel respondents (*i.e.*, those who provide data for all waves for which they are eligible) to compensate for the panel nonrespondents.

In the SIPP, a national probability sample of households is interviewed each year, and all the adults aged 15 and over living in those households at the initial wave become panel members who are followed for the duration

<sup>1</sup> Lou Rizzo, Graham Kalton and J. Michael Brick, Westat Inc., 1650 Research Blvd., Rockville, MD 20850, U.S.A.

of the panel. Until now SIPP panels have had a lifetime of 2 3/4 years, but this is being increased with the 1996 panel to 4 years. Interviews are conducted with panel members at four-month intervals to collect data about income amounts received, participation in income maintenance programs, and other factors that may affect their income and economic welfare. Data are also collected about children. See Nelson, McMillen and Kasprzyk (1985) and Jabine, King and Petroni (1990) for further information on the SIPP design.

The investigation reported here was conducted with the 1987 SIPP panel, using the panel's public use data file. That panel started with a sample of about 12,300 households and followed panel members for seven waves of data collection. The household nonresponse rate at the initial wave was 6.7 percent (Jabine *et al.* 1990). Including children, 30,841 individuals were living in the responding households at the initial wave. Of these individuals, 20.8 percent failed to provide data for all waves for which they were eligible, *i.e.*, they were panel nonrespondents.

In addition to selecting auxiliary variables and studying alternative methods of using those variables to form weighting adjustments for panel nonresponse, this research includes a comparative evaluation of the procedures. The evaluation is performed by comparing a range of estimates produced with the alternative methodologies with one another and with benchmark estimates. The final section of this article summarizes the results and draws conclusions about the effectiveness of the alternative weighting schemes investigated. Further details are given by Rizzo, Kalton, and Brick (1994).

## 2. PREDICTORS OF RESPONSE PROPENSITY

The first step in developing panel nonresponse adjustments is deciding which of the large number of items available from the first wave of data collection should be selected for use in the adjustment procedures. That selection is the focus of this section. The approach adopted is to choose items with responses that discriminate persons by their likelihood to respond at all later waves. Little (1986) calls this method a response propensity stratification method and shows that the large sample bias of estimates can be reduced by adjusting the base weight by the inverse of the probability that an element responds.

In the 1987 SIPP panel, there were 58 items available from the initial wave of data collection (Wave 1) that could be used as potential explanatory variables for panel nonresponse. All of the items used currently by the Bureau of the Census for the SIPP panel nonresponse adjustment were part of this set of 58, with the exception of the Metropolitan Statistical Area (MSA) status, which was suppressed from the public use data file because of disclosure concerns.

With panel response status (panel respondent vs. panel nonrespondent) as the dependent variable, logistic regression analysis was viewed as a natural method for selecting a model for panel nonresponse. However, before attempting this modeling, an initial screening of the variables was performed to reduce the large number of variables to a more manageable set. As a general guideline, items were retained for the logistic regression analysis if the difference in response rates between any two categories for the item was both statistically significant and at least four percentage points. For a variety of reasons, some items were retained even if they did not meet these requirements. For example, the difference in the panel response rates for males and females was less than 2 percent, but gender was nevertheless used in some subsequent analyses.

The screening process reduced the number of items for the logistic regression analysis from 58 to 31. The items retained were: tenure, public housing, household type, Census region, household education, household size, household income, whether householder holds financial instruments (bonds), gender, race, Hispanic origin, relationship to reference person (RRP), age, marital status, family type, education, student status, whether laid off work, personal income, whether holds multiple jobs, working class, whether a recipient of Medicare benefits, Medicaid, Women, Infants, and Children (WIC), Aid to Families with Dependent Children (AFDC), food stamps, general assistance, Social Security, other welfare, Veteran's status, and the number of imputed items at Wave 1.

The last item, the number of imputed items, was included as an index of cooperation at Wave 1. Other studies have found that individuals who are less cooperative at the initial wave of a panel survey are more likely to be nonrespondents at later waves (see, for example, Kalton, Lepkowski, Montanari and Maligalig 1990). As described below, this index turned out to be highly related to panel nonresponse.

### 2.1 Logistic Regression Analysis

Since all 31 items identified in the screening analysis were at least marginally correlated with panel nonresponse, they are all candidate variables for use in a weighting adjustment scheme to reduce the panel nonresponse bias in the survey estimates. However, the screening analysis was limited because it did not consider the interrelationships between the items and it retained too many variables for practical use in making the panel nonresponse adjustments. For example, two items that are highly associated with response status might also be highly correlated with each other, so that the use of one of the two might be sufficient in making the adjustments. To address this issue, the next step in selecting predictors of panel nonresponse was to investigate which combinations of the items could best predict panel response status.



**Table 1**  
Parameter Estimates for the Logistic Regression Model

Predictors	Parameter Estimate
Intercept	-0.465
Age ( $\chi^2 = 184.9$ , $p$ -value < .0001).	
< 16	-0.179
16-24	0.446
25-50	0.187
51-71	-0.056
> 71	0.0
Race ( $\chi^2 = 214.0$ , $p$ -value < .0001).	
White	-0.351
Black	0.255
Other	0.0
RRP ( $\chi^2 = 69.0$ , $p$ -value < .0001).	
Family member	-0.251
Nonfamily member	0.0
Census region ( $\chi^2 = 327.3$ , $p$ -value < .0001).	
New England	0.009
Mid Atlantic	0.167
South Atlantic	0.027
East South Central	-0.231
North Central	-0.396
Mountain/West South Central	0.425
Pacific	0.0
Tenure ( $\chi^2 = 207.2$ , $p$ -value < .0001).	
Home owner	-0.154
Renter	0.331
Other	0.0
Items imputed ( $\chi^2 = 434.2$ , $p$ -value < .0001).	
0	-0.626
1	-0.244
2 to 3	0.296
> 3	0.0
Bond status ( $\chi^2 = 97.1$ , $p$ -value < .0001).	
No bonds	0.168
Some bonds	0.0
Layoff ( $\chi^2 = 33.4$ , $p$ -value < .0001).	
Not laid off	-0.179
Laid off	0.0
Food stamps ( $\chi^2 = 39.3$ , $p$ -value < .0001).	
Not recipient	-0.191
Recipient	0.0
Class of work ( $\chi^2 = 31.4$ , $p$ -value < .0001).	
Business	0.100
Other	0.103
Government	0.0
Education ( $\chi^2 = 12.8$ , $p$ -value = .0003).	
Last grade tenth or eleventh	-0.075
Other	0.0
Household income ( $\chi^2 = 14.9$ , $p$ -value = .0006).	
Less than \$1,200/month	0.117
\$1,200-\$8,000/month	-0.088
Greater than \$8,000/month	0.0
Gender ( $\chi^2 = 10.3$ , $p$ -value = .0013).	
Male	0.047
Female	0.0
RRP-Age < 16 Interaction ( $\chi^2 = 10.1$ , $p$ -value = .0015).	
Family member, child	0.096
Other	0.0

A logistic regression approach was used to examine the joint relationships of several items with panel response status. The regression models were fitted using the Wave 1 survey weights that accounted for unequal selection probabilities and initial wave nonresponse. After examining a number of possible models, a model with thirteen main-effect variables and one interaction term was selected as a reasonable representation of the data.

Table 1 presents the parameter estimates for each level of each predictor variable in this model, together with Wald ( $\chi^2$ ) statistics for each predictor variable. The parameter value of the last level of each predictor variable (the benchmark level) is set to zero. The parameter estimates for the remaining levels of each predictor variable represent differences in response propensity from the benchmark level. As can be seen from the Wald statistics, all the predictor variables make highly significant contributions to the model.

A notable feature of this model is that it contains only one interaction term, the relationship to reference person/age under 16 interaction. All other interactions investigated had smaller  $\chi^2$  values than this one. Even the relationship to reference person/age under 16 interaction has a relatively low predictive power. In fact, this interaction and the last three predictor variables in Table 1 (education, household income, and gender) were not included in most of the weighting procedures discussed below because of their limited predictive power for panel response status. The weighting procedures are mostly based on a reduced main-effects model comprising the first ten predictor variables listed in Table 1.

### 3. ALTERNATIVE WEIGHT ADJUSTMENTS

The method used in the SIPP to adjust the weights for panel nonresponse is described by Chapman, Bailey, and Kasprzyk (1986). The method basically consists of forming nonresponse adjustment cells and then adjusting the weights by the inverses of the response rates in the cells. The cells are formed by the cross-classification of the responses from a set of Wave 1 variables thought to be correlated with panel response. Small cells are combined so that the resulting sample size in each collapsed cell is 30 or more. The reciprocal of the observed (weighted) response rate in each collapsed cell is the panel nonresponse adjustment for that cell. The panel nonresponse adjustment is then multiplied by the Wave 1 weight to create a nonresponse adjusted weight. The Wave 1 weight includes an adjustment for Wave 1 nonresponse, but it does not include the Wave 1 poststratification adjustment.

This section examines alternative methods for performing the panel nonresponse adjustments. These methods can be categorized into three groups:

- Logistic regression methods.
- CHAID methods.
- Generalized raking methods.

Each of the alternative approaches to nonresponse adjustment is discussed below. The procedures for developing the weighting adjustments are detailed along with important statistical properties of the adjustments.

### 3.1 Adjustments Based on Logistic Models

The first set of weighting adjustments we discuss is developed directly from the logistic regression model described in the previous section. This panel nonresponse weighting adjustment, called the *predicted logistic adjustment*, was computed by taking the inverses of the response rates predicted from the reduced main-effects logistic regression model for each of the cells in the crossclassification of the ten predictor variables in that model.

Since the parameters for computing the predicted response rates are estimated with a main-effects model from the marginal responses for the variables, the small sample sizes in the cells of the crossclassification of all the variables are not a concern. However, this benefit is gained by relying completely on the validity of the main-effects model, that is, by assuming that there are no interactions between the variables that need to be taken into account.

One approach to placing less reliance on the main-effects model is to base the adjustments on the observed response rates in cells that have sample sizes large enough to ensure the stability of the observed response rates and to base the adjustments on the predicted response rates in other cells. The second member of the class of alternative adjustments based on logistic regression uses this mixed strategy. In cells containing 25 or more sample persons, the nonresponse adjustment is the inverse of the observed cell response rate. In cells containing less than 25 sample persons, the nonresponse adjustment is the inverse of the predicted response rate for the cell. This adjustment is called the *mixed logistic adjustment*.

A third logistic nonresponse adjustment studied is similar to the current SIPP procedures. Initial cells were defined by the crossclassification of the ten independent variables used in the logistic regression. The cells were then collapsed until the sample size in each cell exceeded 30, and the inverse of the observed response rate within a collapsed cell was then used as the nonresponse adjustment. The strategy for collapsing cells was to group together cells with similar predicted response rates. This nonresponse adjustment is called the *collapsed logistic adjustment*. Although this adjustment is similar to the current SIPP panel nonresponse adjustment, there are some differences in the variables used to define the cells and the methods used to combine small cells are different.

For all three alternative weighting adjustments based on the logistic regression model, the observed and predicted response rates were computed from weighted counts of the number of cases rather than using the unweighted numbers, where the weights were the nonresponse adjusted Wave 1 weights. In practice, the weighted and unweighted adjustments were nearly the same.

#### 3.1.1 Adjustments Based on CHAID Models

The second class of methods for adjusting for panel nonresponse involved using the CHAID categorical search algorithm to divide the data set into adjustment cells. The general approach was to define adjustment cells as combinations of responses to the predictor variables that had the greatest discrimination with respect to panel response rates, subject to the restriction that each cell should have a minimum sample size of at least 25 persons. The panel nonresponse adjustment was the inverse of the observed response rate in the cell.

The CHAID algorithm creates cells by splitting the data set progressively in a tree structure. The splitting along each newly created branch is performed by choosing the variable that maximizes a  $\chi^2$  criterion. When the split involves a polychotomous variable, the split may involve several branches. The  $\chi^2$  tests are modified using Bonferroni type adjustments to prevent variables from being chosen simply because they have more categories. CHAID is one version of the Automatic Interaction Detector (AID) developed for categorical variables. Kass (1980) presents the theory underlying the CHAID technique. Another version of the same methodology was used by Lepkowski, Kalton and Kasprzyk (1989) and Kalton, Lepkowski and Lin (1985) to model nonresponse in SIPP.

For the current analysis, two CHAID models were examined by including different sets of predictor variables. The first model included the seven most important predictors in the logistic regression model (age, relationship to reference person, race of householder, tenure, Census region, imputation flags, and bond-holding status), plus gender. This model resulted in 99 nonresponse adjustment cells. The nonresponse adjustment based on this model is called CHAID 1. The second CHAID model included the 13 predictor variables from the logistic regression model presented in Table 1. This model resulted in 142 nonresponse adjustment cells. The nonresponse adjustment for this model is called CHAID 2.

#### 3.1.2 Adjustments Based on Generalized Raking

The third class of methods examined for adjusting for panel nonresponse was generalized raking. Unlike the other approaches, nonresponse adjustment cells were not developed by crossclassifying the predictor variables. Rather, raking was directly applied to force the panel



respondents' marginal distributions for each of the predictor variables (computed using the adjusted weights) to equal the corresponding distributions for respondents and nonrespondents combined (computed using the original Wave 1 weights). Kalton and Kasprzyk (1986) refer to this method as sample based raking. The ten predictor variables from the reduced logistic regression model were used to define the marginal distributions. Hence, the raking problem was ten dimensional, with one dimension for each predictor variable.

Raking involves modifying the original weights in order to satisfy certain marginal constraints while minimizing the distance between the original and adjusted weights. Deville and Särndal (1992) describe some distance functions that may be used and derive the corresponding raking methodologies. The raking algorithm of Deming and Stephan (1942), which implicitly employs a distance function that leads to a multiplicative solution, is one form of generalized raking.

The CALMAR software described by Deville, Särndal and Sautory (1993) was used to compute the adjustments. Three different distance functions were examined: the multiplicative method, the linear method, and the truncated multiplicative method. The adjustments for all three distance functions were found to be nearly identical. This empirical result is consistent with results given by Deville and Särndal (1992) that show that the estimators using weights generated with different distance functions are asymptotically equivalent if the distance functions satisfy certain smoothness conditions. The three distance functions employed in this research satisfy those conditions. Since the adjustments were nearly identical for all three methods, only the weighting adjustment from the multiplicative method was retained for further evaluation. The resulting adjustment is called the *raking* adjustment.

### 3.1.3 Distributions of Nonresponse Adjustments

The adjustments for each of the six schemes described above were computed for the 1987 SIPP panel file. Table 2 summarizes the distributions of the resulting nonresponse adjustments. The summary is for the adjustments only, not the weights that are the products of the adjustments and the Wave 1 weights. Table 2 is divided into two parts: the upper part shows the mean, median, and extreme values for each adjustment distribution, as well as  $(1 + CV^2)$ , where CV is the coefficient of variation for each adjustment. The statistic  $(1 + CV^2)$  serves as an indicator of the increase in variance of the estimates introduced by having variable nonresponse adjustment factors (see Kish 1992). The second part of Table 2 shows the correlations among the alternative forms of adjustment.

Since the overall weighted panel response rate is 0.794, the mean overall nonresponse adjustment would be  $1/(0.794) = 1.26$  if the same adjustment were used for all persons. The mean weighting adjustments for the three weighting adjustments that use the inverses of cell response rates (collapsed logistic, CHAID 1 and CHAID 2) are necessarily equal to the overall nonresponse adjustment of 1.26. The mean weighting adjustments for the other schemes differ only minimally from the mean overall nonresponse adjustment.

For all six schemes, the distributions are positively skewed, with a few cases with large weights. By their nature, the various logistic and CHAID schemes cannot have adjustments less than 1.00, whereas the raking algorithm can, and does, do so. The median weights are similar among all schemes, but the maximum weights are not. The CHAID 2 scheme has a cell with a response rate of only 7 percent, leading to the largest maximum weight of 13.93. The raking scheme has the smallest maximum weight of 2.51.

**Table 2**  
Distribution of Panel Nonresponse Adjustments

	Mean	Minimum	Median	Maximum	1 + CV <sup>2</sup>	
Predicted logistic	1.26	1.04	1.20	4.28	1.02	
Mixed logistic	1.26	1.00	1.20	4.28	1.03	
Collapsed logistic	1.26	1.00	1.20	3.43	1.02	
CHAID 1	1.26	1.02	1.22	3.49	1.03	
CHAID 2	1.26	1.01	1.19	13.93	1.04	
Raking	1.26	0.91	1.23	2.51	1.02	
Correlations						
	Predicted Logistic	Mixed Logistic	Collapsed Logistic	CHAID 1	CHAID 2	Raking
Predicted logistic	1.00	0.96	0.73	0.73	0.63	0.95
Mixed logistic		1.00	0.73	0.72	0.63	0.90
Collapsed logistic			1.00	0.69	0.58	0.75
CHAID 1				1.00	0.81	0.73
CHAID 2					1.00	0.63
Raking						1.00

The values of  $(1 + CV^2)$  are fairly consistent across the various adjustments. The CHAID 2 adjustment has the greatest value of  $(1 + CV^2)$ , primarily because of the presence of more outlying adjustments (such as the maximum value of 13.93). However, even for this method, the approximate increase in the variance of the survey estimates is only four percent. The raking adjustment has the smallest increase in variance (two percent), but this increase is not very different from that of the other methods.

The pairwise correlations between the six alternative sets of weights range from 0.58 to 0.96. Not surprisingly, the predicted logistic and mixed logistic weights are highly correlated. Given the similarity of the predicted main-effects logistic regression scheme to raking, it is also not surprising that their two sets of weights are highly correlated. The relatively high correlation between the raking weights and the CHAID 1 weight and the collapsed logistic weight is consistent with the earlier result showing no large interaction terms. The CHAID 2 weights have the lowest correlations with the other sets of weights, except for their correlation with the CHAID 1 weights. This finding is probably explained by the wide variability in the CHAID 2 weights resulting from the use of as many as 142 adjustment cells.

### 3.2 Final Panel Weights

The panel nonresponse adjustment weights discussed in the previous section represent the adjustments to the Wave 1 weights to compensate for panel nonresponse. The final panel weights that may be used in the analysis of the SIPP panel file are obtained by multiplying the panel nonresponse adjustment weights by the Wave 1 weights, and then applying poststratification to make weighted sample totals conform to totals derived primarily from the Current Population Survey (CPS). This procedure was applied for each of the six alternative panel nonresponse adjustment schemes.

The poststratification procedure used was equivalent to the current SIPP procedure, except that the latter procedure poststratifies by rotation groups whereas for the alternative weighting schemes the poststratification was performed on all rotation groups combined. The difference should not have an appreciable effect. After poststratification, the six alternative sets of final weights and the SIPP panel weights sum to the same control totals.

To compare the final panel weights for the six adjustment schemes with one another and with the current SIPP panel weight, the correlations between the weights were computed, along with the measure of variability used previously,  $(1 + CV^2)$ . The results are presented in Table 3. The estimates of the variability due to the weighting  $(1 + CV^2)$  indicate similar increases of between 8 and 10 percent in the variances of survey estimates for all of the weighting schemes. The correlations between the alternative sets of final panel weights are all 0.85 or higher. Comparing these correlations to those in Table 2, it is clear that the correlations between the final weights are appreciably higher than those between the panel nonresponse adjustment weights. The correlations between the SIPP panel weight and the alternative final weights are consistently lower than any others, probably because the variables used in forming the nonresponse adjustments for this weight differed from those used for the alternative weights. The variables used in the alternative schemes that are not used in the SIPP panel weight are age, relationship to reference person, number of imputed items, class of work, and food stamp reciprocity. Household size is the only variable other than MSA status (which was not available due to disclosure concerns) used in the SIPP panel weight but not used for the alternative schemes because it was not found to be significantly associated with response rates.

**Table 3**  
Correlations Between Poststratified Weights with Variance Inflation Measures

	SIPP panel	Predicted Logistic	Mixed Logistic	Collapsed Logistic	CHAID 1	CHAID 2	Raking
SIPP panel	1.00	0.75	0.74	0.75	0.71	0.68	0.77
Predicted logistic		1.00	0.99	0.91	0.90	0.86	0.98
Mixed logistic			1.00	0.91	0.90	0.86	0.97
Collapsed logistic				1.00	0.89	0.85	0.93
CHAID 1					1.00	0.94	0.91
CHAID 2						1.00	0.87
Raking							1.00
$1 + CV^2$	1.08	1.09	1.09	1.08	1.09	1.10	1.08



#### 4. COMPARING ESTIMATES USING ALTERNATIVE WEIGHTS

The previous section described the development of the alternative sets of final weights that may be used for the analysis of the SIPP panel file. All the final weighting schemes incorporate adjustments for unequal selection probabilities, nonresponse at the initial wave, panel nonresponse, and poststratification to external control totals. This section compares survey estimates obtained using the alternative weighting schemes with one another and with the corresponding estimates obtained using the SIPP panel weights. In addition, where possible, the various survey estimates are also compared with external estimates from other sources. Some of the external estimates are benchmark estimates obtained from administrative records or the Current Population Survey. Other external estimates are obtained from Wave 1 of the 1989 SIPP panel. Data collected in Wave 7 of the 1987 SIPP panel relate to the same time period as data collected in Wave 1 of the 1989 SIPP panel, and hence estimates obtained from these two data sources should be comparable.

In making comparisons with benchmark estimates, it needs to be recognized any differences observed may be explained by a variety of factors of which panel nonresponse is only one. For example, response errors and differences in definitions may explain differences between SIPP estimates and benchmark estimates. Thus the benchmark comparisons need to be treated with caution. Since the 1989 SIPP panel estimates are based on Wave 1 data, they are not subject to the panel nonresponse. Thus, differences between estimates obtained from the 1987 and 1989 SIPP panels are perhaps the most likely to be caused by a failure of the panel nonresponse adjustments to fully compensate for panel nonresponse bias. However, even in this case, alternative explanations such as panel conditioning could contribute to the differences (although Pennell and Lepkowski 1992, show that panel conditioning is not a major factor in most SIPP estimates).

Table 4 presents a variety of estimates from the 1987 SIPP panel file using the SIPP panel weight and the six alternative weighting schemes, and corresponding benchmark estimates and estimates from the 1989 SIPP panel where available. The estimates are percentages, except for the estimates of the mean number of months without health insurance, median household income, and annual wages. The estimates are for the total population, except for the employment estimates (percent employed, unemployed and out of the labor force), which are for persons over the age of 15, and for annual wages, which are for persons over the age of 14. The estimates are for three different time periods: June 1987, January 1989, and the calendar year of 1987. For example, the first three estimates in Table 4 are the estimated percentages of

persons participating in the AFDC (Aid for Families with Dependent Children) program in June 1987, in January 1989, and at any time during the 1987 calendar year. A comparable estimate from the 1989 SIPP panel is available only for the January 1989 time period.

The most notable finding from Table 4 is the similarity of the estimates computed with all the weighting schemes from the 1987 panel. The percentage estimates in Table 4 are in fact given to two decimal places because the use of the conventional one decimal place would often show no difference between the alternative estimates. The largest difference occurs for the percentage employed in January 1989, where the estimate using the SIPP panel weight is 62.7 percent and the estimate using the mixed logistic regression weight is 62.3 percent. Even this largest of differences is relatively small, especially when considering that the estimated standard error for this estimate is 0.3 percent.

When the 1987 SIPP panel estimates are compared with the external estimates from the 1989 SIPP panel and from other sources, some of the differences are much larger and of substantive importance. To examine these differences in more detail, standardized differences between the alternative estimates and the benchmark estimates were computed and are shown in Table 5. A standardized difference is defined as the difference between the alternative estimate and the external estimate divided by the standard error of the difference.

The upper part of Table 5 shows the standardized differences when the 1989 SIPP panel is used to produce the external estimate. The standardized differences for most of the estimates are less than 2.0 in absolute value, indicating that the differences may be accounted for by sampling error. However, the standardized differences for the percentage unemployed and for the poverty rate are greater than 2.0 and highly significant. Thus, the alternative weighting adjustments do not succeed in bringing the 1987 survey estimates in line with the 1989 survey estimates for all characteristics.

The lower part of Table 5 shows the standardized differences when other benchmark estimates are used. These standardized differences are generally large and in many cases very large. Only a few are less than 2.0 and many are greater than 10.0. Given the much smaller standardized differences found in the upper part of Table 5 for similar statistics, it seems likely that factors other than panel nonresponse bias are largely responsible for the magnitude of these differences. The standardized differences based on these largely administrative data sources may signal important issues related to the quality of the data (from either the SIPP, the benchmark data source, or both), but they do not provide much help in assessing the effectiveness of alternative nonresponse adjustments in reducing panel nonresponse bias.

**Table 4**  
Estimates for the Total Population from the 1987 SIPP Panel with Alternative Weighting Schemes  
and Estimates from Other Sources

	SIPP Panel	Predicted Logistic	Mixed Logistic	Collapsed Logistic	CHAID 1	CHAID 2	Raking	1989 SIPP Panel	Bench- mark
AFDC – June 1987	3.73	3.70	3.74	3.72	3.71	3.60	3.69		4.28 <sup>1</sup>
AFDC – January 1989	3.10	3.12	3.14	3.12	3.14	3.02	3.10	3.56	4.24 <sup>2</sup>
AFDC – Annual 1987	4.85	4.78	4.82	4.81	4.80	4.69	4.78		
Food stamps – June 1987	7.43	7.26	7.30	7.34	7.38	7.20	7.21		7.35 <sup>3</sup>
Food stamps – January 1989	6.71	6.63	6.67	6.64	6.70	6.59	6.58	6.30	7.29 <sup>4</sup>
Food stamps – Annual 1987	10.30	10.11	10.16	10.18	10.24	10.05	10.06		
Medicaid – January 1989	6.77	6.78	6.81	6.75	6.81	6.68	6.76	6.97	
Medicaid – Annual 1987	9.21	9.21	9.24	9.21	9.25	9.09	9.21		
SSI – June 1987	1.68	1.70	1.69	1.67	1.69	1.65	1.69		1.68 <sup>3</sup>
SSI – January 1989	1.65	1.67	1.66	1.64	1.66	1.61	1.66	1.65	1.74 <sup>3</sup>
SSI – Annual 1987	1.80	1.82	1.82	1.80	1.82	1.78	1.82		
Social security – January 1989	14.92	14.87	14.87	14.89	14.88	14.89	14.85	15.14	
Poverty rate – June 1987	10.88	10.75	10.79	10.76	10.79	10.69	10.74		
Poverty rate – January 1989	12.91	12.98	13.02	12.97	12.99	12.91	12.93	14.46	
Entering poverty 1987/1988	2.25	2.31	2.32	2.30	2.29	2.32	2.31		
Leaving poverty 1987/1988	2.69	2.63	2.64	2.60	2.62	2.63	2.63		
Mean months without health insurance – 1987	1.66	1.69	1.70	1.67	1.67	1.69	1.69		
Median household income – January 1989	2,601	2,600	2,597	2,607	2,607	2,607	2,602	2,550	
Annual wages 1987 (in trillions)	1.93	1.94	1.93	1.94	1.94	1.94	1.94		2.22 <sup>4</sup>
Employed – January 1989	62.74	62.36	62.34	62.43	62.42	62.52	62.42	61.60	
Unemployed – January 1989	3.57	3.64	3.63	3.60	3.58	3.60	3.63	4.52	
Out of labor force – January 1989	33.69	34.01	34.03	33.96	34.01	33.88	33.95	33.88	
Married in 1987	1.39	1.41	1.40	1.39	1.39	1.39	1.41		1.86 <sup>5</sup>
Divorced in 1987	0.51	0.50	0.50	0.49	0.50	0.51	0.49		0.90 <sup>6</sup>
Changed address in 1987	12.88	13.32	13.32	13.19	13.36	13.37	13.33		17.99 <sup>6</sup>

<sup>1</sup> Social Security Bulletin, Volume 52, No. 3.<sup>2</sup> Social Security Bulletin, Volume 51, No. 7.<sup>3</sup> USDA Food and Nutrition Service, unpublished data.<sup>4</sup> U.S. Bureau of the Census, Current Population Reports, Consumer Income, P-60, No. 174.<sup>5</sup> National Center for Health Statistics: Vital Statistics of the U.S., 1987, Volume III, Marriage and Divorce, DHHS Pub. No. (PHS) 91-1103.<sup>6</sup> U.S. Bureau of the Census, Current Population Reports, Population Characteristics, P-20, No. 473.



**Table 5**  
Standardized Differences Between 1987 SIPP Panel Estimates and Benchmark Estimates

	Benchmark Estimate	SIPP Panel	Predicted Logistic	Mixed Logistic	Collapsed Logistic	CHAID 1	CHAID 2	Raking
<b>1989 SIPP panel estimates</b>								
AFDC	3.56	-1.58	-1.52	-1.43	-1.52	-1.44	-1.84	-1.57
Food stamps	6.30	1.02	0.82	0.92	0.86	1.01	0.73	0.69
Medicaid	6.97	-0.50	-0.47	-0.40	-0.53	-0.39	-0.70	-0.51
SSI	1.65	0.05	0.11	0.08	-0.03	0.07	-0.15	0.09
Social Security	15.14	-0.38	-0.46	-0.46	-0.42	-0.44	-0.42	-0.50
Poverty rate	14.46	-2.77	-2.64	-2.57	-2.67	-2.63	-2.78	-2.74
Median Income	2,550	2.05	2.01	1.89	2.30	2.30	2.29	2.09
Employed	61.60	2.42	1.60	1.56	1.76	1.72	1.95	1.73
Unemployed	4.52	-4.93	-4.59	-4.59	-4.76	-4.90	-4.78	-4.60
Out of labor force	33.88	-0.42	0.28	0.32	0.18	0.28	-0.01	0.15
<b>Other benchmark estimates</b>								
AFDC - June 1987	4.28	-2.55	-2.66	-2.49	-2.59	-2.65	-3.14	-2.71
AFDC - January 1989	4.24	-5.71	-5.62	-5.49	-5.63	-5.51	-6.10	-5.70
Food stamps - June 1987	7.35	0.27	-0.31	-0.16	-0.04	0.11	-0.50	-0.48
Food stamps - January 1989	7.29	-2.04	-2.32	-2.17	-2.26	-2.06	-2.44	-2.50
SSI - June 1987	1.68	0.00	0.13	0.08	-0.03	0.08	-0.20	0.11
SSI - January 1989	1.74	-0.57	-0.48	-0.53	-0.67	-0.54	-0.84	-0.50
Annual wages 1987	2.22	-16.12	-15.94	-16.38	-15.66	-15.61	-15.60	-15.78
Married in 1987	1.86	-5.11	-4.93	-4.98	-5.11	-5.10	-5.07	-4.95
Divorced in 1987	0.90	-7.15	-7.37	-7.36	-7.40	-7.32	-7.20	-7.40
Changed address in 1987	17.99	-11.49	-10.50	-10.51	-10.80	-10.42	-10.40	-10.49

## 5. DISCUSSION

Nonresponse weights are widely used to compensate for unit nonresponse in sample surveys. The basic requirement for this form of weighting is the availability of information on one or more auxiliary variables for both respondents and nonrespondents. In many surveys, this information is available for only a small number of auxiliary variables (such as the PSUs and strata from which the units were selected). In such surveys, the nonresponse weights can often be simply developed as weighting class adjustments for a set of classes based on the crosstabulation of the auxiliary variables.

There are, however, surveys in which data are available for a large number of auxiliary variables for possible use in developing nonresponse weights. This situation often applies when an administrative record system is used as the survey's sampling frame, with all the information in the system then being available for use in making nonresponse adjustments. It also applies when the survey data collection is conducted in two or more phases (*e.g.*, an initial screening interview followed by a detailed interview or some other form of data collection at a later time point) and when nonresponse adjustments are needed for later

phases; in this case, data from prior phases of data collection may be used in compensating for nonresponse at later phases. A similar situation applies in panel surveys when adjustments are required for nonresponse at later waves of the panel, as discussed in this paper.

When a large number of auxiliary variables is available for all sampled units, two main choices need to be made. First, there is the choice of auxiliary variables to use in the adjustment. Second, there is the choice of the adjustment method to be applied.

The basic approach adopted in this study for choosing the auxiliary variables for use in the nonresponse adjustment was to identify the set of variables that were good predictors of panel nonresponse. With so many auxiliary variables available, the first step was a screening procedure to eliminate variables that were found to have little association with the panel nonresponse rate. Then, logistic regression models using predictor variables remaining from the screening were examined to identify the set of variables to be retained for use in adjusting the weights. Whether the number of auxiliary variables is reduced to a manageable set by this or some other approach (*e.g.*, by using the CHAID algorithm), this reduction is likely to be a necessary first step when there are many potential auxiliary variables available.

After selecting the subset of auxiliary variables, a wide variety of methods exists for creating the nonresponse adjustments. We examined panel nonresponse adjustments based on logistic regression models, categorical search models, and sample-based generalized raking. The final panel weights resulting from these adjustment schemes were highly correlated with one another and they yielded estimates that were very similar. None of the schemes produced estimates that were superior in terms of bias reduction.

In part, the high correlation of the final panel weights generated by the different adjustment schemes may be explained by the similarity of many of the adjustment schemes. In part, it may be explained by the final post-stratification weighting which raised the correlations between the weights. It may also be partly explained by the lack of large interaction effects between the auxiliary variables. If there were sizable interaction effects that were not included in the logistic modeling, then one might expect greater differences between the raking and predicted logistic weights on the one hand and the CHAID, mixed logistic, and collapsed logistic weights on the other hand. Thus, the similarity in weights produced by the alternative weighting schemes for the SIPP may not be as great in other circumstances.

A common concern that arises when many auxiliary variables are used to adjust the weights is that the adjusted weights might be highly variable, thus causing a serious loss of precision in the survey estimates. This proved not to be the case in the methods we evaluated. The variability of the weights with all the weighting schemes turned out to be similar, provided reasonable precautions were taken in creating the adjustments.

Although the empirical results do not show any appreciable differences in the estimates produced using the alternative weighting schemes and those produced using the SIPP panel weights, the correlations of the alternative adjusted weights and the current SIPP panel weight were found to be lower than the correlations among the alternative weights. This finding suggests that the choice of auxiliary variables is an important one, and probably more important than the choice of the weighting methodology. Although the more systematic methods used in this research for choosing the auxiliary variables did not result in major improvements over the current SIPP procedures, an analytic based choice of auxiliary variables may be more productive in other studies.

When a sizable number of auxiliary variables that are correlated to response propensity is available, it seems wise to use as many of them as possible in the nonresponse adjustment to serve as a safeguard in attempting to compensate for nonresponse bias. This general strategy should, however, be tempered by a careful assessment of the variation of the resulting weights in order to avoid too great a loss of precision in the survey estimates. In addition,

a practical consideration that should be taken into account is the ease of implementation of the weighting methodology. If, as in this study, alternative weighting methodologies yield very similar weights and estimates, a method that is simple to apply may be preferable.

## ACKNOWLEDGMENTS

We thank the referees for helpful comments on the earlier version of the paper. The paper reports research undertaken for the U.S. Bureau of the Census. The views expressed are the authors'. They do not necessarily reflect the views of the Bureau of the Census.

## REFERENCES

- CHAPMAN, D.W., BAILEY, L., and KASPRZYK, D. (1986). Nonresponse adjustment procedures at the U.S. Bureau of the Census. *Survey Methodology*, 12, 161-180.
- DEMING, W.E., and STEPHAN, F.F. (1942). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11, 427-444.
- DEVILLE, J.-C., and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- DEVILLE, J.-C., SÄRNDAL, C.-E., and SAUTORY, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.
- JABINE, T.B., KING, K.E., and PETRONI, R.J. (1990). *Survey of Income and Program Participation (SIPP): Quality Profile*. Washington, DC: U.S. Bureau of the Census.
- KALTON, G. (1983). *Compensating for Missing Survey Data*. Ann Arbor, MI: Survey Research Center, University of Michigan.
- KALTON, G., and KASPRZYK, D. (1986). The treatment of missing survey data. *Survey Methodology*, 12, 1-16.
- KALTON, G., LEPKOWSKI, J.M., MONTANARI, G.E., and MALIGALIG, D. (1990). Characteristics of second wave nonrespondents in a panel survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 462-467.
- KALTON, G., LEPKOWSKI, J., and LIN, T. (1985). Compensating for wave nonresponse in the 1979 ISDP Research Panel. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 372-377.
- KASS, G.V. (1980). An exploratory technique for investigating large quantities of categorical data, *Applied Statistics*, 29, 119-127.
- KISH, L. (1992). Weighting for unequal  $P_i$ . *Journal of Official Statistics*, 8, 183-200.



- LEPKOWSKI, J., KALTON, G., and KASPRZYK, D. (1989). Weighting adjustments for partial nonresponse in the 1984 SIPP Panel. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 296-301.
- LITTLE, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*, 54, 2, 137-139.
- NELSON, D., McMILLEN, D., and KASPRZYK, D. (1985). *An Overview of the SIPP, Update 1*. SIPP Working Paper No. 8401. Washington, DC: U.S. Bureau of the Census.
- OH, H.L., and SCHEUREN, F. (1983). Weighting adjustments for unit nonresponse. In *Incomplete Data in Sample Surveys, Volume 2: Theory and Bibliographies* (Eds. W.G. Madow, I. Olkin, and D. Rubin), 143-184. New York: Academic Press.
- PENNELL, S.G., and LEPKOWSKI, J.M. (1992). Panel conditioning effects in the Survey of Income and Program Participation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 556-571.
- RIZZO, L., KALTON, G., and BRICK, M. (1994). Weighting adjustments for panel nonresponse in the SIPP. Final Report submitted to U.S. Bureau of the Census for SIPP Panel Nonresponse Project.





# Multiple Sample Estimation of Population and Census Undercount in the Presence of Matching Errors

YE DING and STEPHEN E. FIENBERG<sup>1</sup>

## ABSTRACT

The multiple capture-recapture census is reconsidered by relaxing the traditional perfect matching assumption. We propose matching error models to characterize error-prone matching mechanisms. The observed data take the form of an incomplete  $2^k$  contingency table with one missing cell and follow a multinomial distribution. We develop a procedure for the estimation of the population size. Our approach applies to both standard log-linear models for contingency tables and log-linear models for heterogeneity of catchability. We illustrate the method and estimation using a 1988 dress rehearsal study for the 1990 census conducted by the U.S. Bureau of the Census.

KEY WORDS: Capture-recapture census; Estimates for total population size; Log-linear models; Matching errors; Multiple recapture census.

## 1. INTRODUCTION

The multiple recapture census technique has been used in many fields to estimate the size of a closed population. Cormack (1968) and Seber (1982) give excellent reviews of many techniques used. Here we consider a sequence of samples,  $s_1, \dots, s_k$ , where the members of  $i$ -th sample are uniquely labeled, for example, by tagging or marking, and then returned to the population (Darroch 1958). Usual multiple recapture census methods make the following assumptions.

- (1) **Perfect matching.** Individuals in one list (information source, sample) can be matched with those in another list without error. In other words, there are no misclassification errors with respect to determining whether a particular individual has been recorded by both information sources or only one of them.
- (2) **Independence.** The lists are independent of one another, that is, the probability of an individual being included in one list does not depend on whether the individual was included in previous lists.
- (3) **Homogeneity (Equal Catchability).** All individuals in the population under study have equal probabilities of being observed (captured) in any list (sample).
- (4) **Closure.** The population in question is "closed", so that there are no changes due to birth, death, emigration, or immigration during the period when the sampling takes place.

Darroch (1958) examined the multiple recapture census under these four assumptions. Fienberg (1972) adopted a log-linear model approach to allow for statistical dependence of specific types among samples, thereby dropping the independence assumption. Darroch, Fienberg, Glonek and Junker (1993) developed an extended log-linear model

approach that allows for individual-level heterogeneity as well as dependence, but it requires at least three samples, *i.e.*,  $k = 3$ . In the context of the two-sample census approach used by U.S. Bureau of Census for census coverage evaluation, matching problems due to unavoidable mismatches and erroneous nonmatches have been explored by several authors. For example, Ding and Fienberg (1994) considered modeling matching errors in the two-sample census and developed systematic procedure for the estimation of population totals. The inclusion of a third sample, *e.g.*, drawn from the administrative records, in modeling and estimation of census coverage has been considered by the U.S. Bureau of Census in the past and remains an option to augment and evaluate the dual system approach. In this paper, we consider matching error models for the multiple sample census problem, allowing for both dependence and heterogeneity.

Here we view the observations from a multiple recapture census data as falling into a  $2^k$  cross-classification, with absence or presence on the  $i$ -th sample defining the category for the  $i$ -th dimension. In this cross-classification, the cell corresponding to absence for all  $k$  samples is missing. The objective is to estimate the number of individuals in the population who are not observed, which corresponds to the missing cell in the  $2^k$  incomplete contingency table. In Section 2, we investigate the effects of matching errors on the observed  $2^k$  incomplete table. In Section 3, some models for matching errors are proposed to characterize an error-prone matching process. Based on these models and assumptions (3) and (4), we develop a procedure using log-linear model formulation for the estimation of the population size. In Section 5, we use the proposed methods to analyze data from 1988 Dress Rehearsal Census conducted by the U.S. Bureau of Census.

<sup>1</sup> Ye Ding, Research Scientist, Bureau of Biometrics, New York State Health Department, Concourse, Room C-144, Empire State Plaza, Albany, New York 12237, U.S.A.; Stephen E. Fienberg, Maurice Falk Professor of Statistics and Social Science, Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, U.S.A.

## 2. MATCHING ERRORS IN MULTIPLE SAMPLE CENSUS

We begin by classifying matching errors into two broad categories, mismatches and erroneous nonmatches. To understand the nature of matching errors in multiple-sample census, we review the case of a three-sample census. Suppose that there are no missing data or errors in recording the information for any individual in the population and one takes three samples from the population,  $s_1$ ,  $s_2$ , and  $s_3$ . For instance, suppose that, in sample  $s_1$ , individuals 1, 3, 4 and 7 are seen, individuals 3, 4, and 8 are seen in  $s_2$ , and individuals 4, 9, and 10 in  $s_3$ . In vector notation, we can represent this as  $s_1 = (1, 3, 4, 7)$ ,  $s_2 = (3, 4, 8)$  and  $s_3 = (4, 9, 10)$ . Matching errors are not present provided that there is complete and correct information available. We thus have the following incomplete  $2^3$  table corresponding to these three samples:

**Table 1**  
Original Table without Matching Errors

$s_3$	$s_1$			
	Present		Absent	
	$s_2$		$s_2$	
	Present	Absent	Present	Absent
Present	1	0	0	2
Absent	1	2	1	-

Suppose further that, because of missing data or incorrect information, we actually observe

$$s_1 = (1, 3, 4, 7), \quad s_2 = (3^*, 4^*, 8), \quad s_3 = (4, 9, 10),$$

where  $3^*$  and  $4^*$  are individuals 3 and 4 but with incorrect information leading to two erroneous nonmatches when the samples are matched. Assuming no erroneous matches, we then observe the incomplete  $2^3$  table:

**Table 2**  
Observed Table with Matching Errors

$s_3$	$s_1$			
	Present		Absent	
	$s_2$		$s_2$	
	Present	Absent	Present	Absent
Present	0	1	0	2
Absent	0	3	3	-

The effects of matching errors are obvious from a comparison of Table 1 and 2:

- (i) The number of observations may increase for some cells while decreasing for the others, and as a consequence, the marginal totals and especially the total number of different individuals observed in the three samples may change, subject to the constraint that the total number of observations in each sample,  $x_{1++}$ ,  $x_{+1+}$ , and  $x_{++1}$ , remain the same. Changes in the total number of different individuals in all samples make our problem distinct from the usual misclassification problem in the analysis of categorical data, in which the possibility of making mistakes in classifying individuals into respective categories is considered. (e.g., see Chen 1979).
- (ii) In parallel, there may be changes in some cell probabilities subject to the constraint that the probability of being captured in a sample,  $p_{1++}$ ,  $p_{+1+}$ , and  $p_{++1}$ , is unchanged.

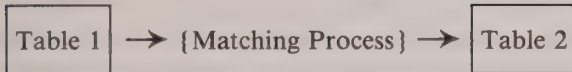
Because of the complexity of matching errors in the three-sample case, we need some special terminology for descriptive convenience. We say that an individual is at state 1 with respect to sample  $s_1$  if the individual is observed in  $s_1$  and at state 0 if not. We use a triple  $(i, j, k)$ ,  $0 \leq i, j, k \leq 1$ , to denote an individual at state  $i, j$ , and  $k$  with respect to  $s_1, s_2$  and  $s_3$ , respectively. For instance,  $(1, 0, 0)$  is an individual observed only in  $s_1$ , and  $(1, 1, 1)$  is an individual captured in three samples. We define the level of an individual  $(i, j, k)$  as  $i + j + k$ , i.e., the number of samples in which the individual is included. There are four different levels, 0, 1, 2 and 3. An individual has level 0 if and only if he/she is not captured by any sample, and has level 3 if he/she is in three samples. For a  $(1, 1, 0)$  individual, if the correct match is not made according to the matching rule, this individual decomposes into "two different" individuals, a  $(1, 0, 0)$  and a  $(0, 1, 0)$ , assuming no erroneous matches. On the other hand, a  $(1, 0, 0)$  individual matched incorrectly with a  $(0, 1, 0)$  will produce a single observed  $(1, 1, 0)$  individual. For convenience, we call such a decomposition or combination a *transition*. Then transitions can only go from level 3 or 2 to the same (if there is no matching error) or lower levels in the absence of erroneous matches. More specifically, a  $(1, 1, 1)$  person may make a transition into one of 5 possible sets of individuals

$$\{(1, 1, 1)\}, \quad \{(1, 0, 0), (0, 1, 1)\}, \quad \{(0, 1, 0), (1, 0, 1)\}$$

$$\{(0, 0, 1), (1, 1, 0)\}, \quad \{(1, 0, 0), (0, 1, 0), (0, 0, 1)\}.$$

For level 2 individuals,  $(1, 1, 0)$  can decompose into  $\{(1, 0, 0), (0, 1, 0)\}$  or stay at  $\{(1, 1, 0)\}$ , and similarly for  $\{(0, 1, 1)\}$  and  $\{(1, 0, 1)\}$ . From above discussions, we summarize the effect of matching errors by the following diagram:





where Table 1 is the original  $2^k$  incomplete table with no matching errors and Table 2 is the observed  $2^k$  incomplete table in the presence of matching errors. Henceforth, we denote the cell probabilities and expected cell counts associated with Table 1 by  $\{r_{ijk}\}$  and  $\{l_{ijk}\}$  and those of Table 2 by  $\{p_{ijk}\}$ ,  $\{m_{ijk}\}$ , for  $1 \leq i, j, k \leq 2$ .

### 3. SOME MODELS FOR MATCHING ERRORS

We now propose models to describe the matching errors, each of which allows us to formulate the reallocation of cell probabilities and expected cell counts associated with Table 1.

**Model (1).** In addition to the homogeneity and closure assumptions in §1, we assume that: (i) There are no erroneous matches in the matching process; (ii) Any individual will stay at his original state with probability  $\theta$ , and transition to any of a possible set of individuals with probability  $(1 - \theta)/(m - 1)$ , where  $m$  is the number of all possible sets of individuals to which the individual may transition. For example, for a (1,1,1) person discussed late in last section,  $m = 5$ .

Under this model, for the three-sample census, we can express the probabilities for the table with matching errors,  $\{p_{ijk}\}$ , in terms of probabilities of the table with no matching errors,  $\{r_{ijk}\}$ :

$$p_{111} = \theta r_{111},$$

$$p_{112} = \frac{1 - \theta}{4} r_{111} + \theta r_{112},$$

$$p_{121} = \frac{1 - \theta}{4} r_{111} + \theta r_{121},$$

$$p_{211} = \frac{1 - \theta}{4} r_{111} + \theta r_{211},$$

$$p_{122} = \frac{1 - \theta}{2} r_{111} + (1 - \theta)r_{112} + (1 - \theta)r_{121} + r_{122},$$

$$p_{212} = \frac{1 - \theta}{2} r_{111} + (1 - \theta)r_{112} + (1 - \theta)r_{211} + r_{212},$$

$$p_{221} = \frac{1 - \theta}{2} r_{111} + (1 - \theta)r_{211} + (1 - \theta)r_{121} + r_{221}.$$

Let

$$\vec{p} = (p_{111}, p_{112}, p_{121}, p_{211}, p_{122}, p_{212}, p_{221})^T,$$

and

$$\vec{r} = (r_{111}, r_{112}, r_{121}, r_{211}, r_{122}, r_{212}, r_{221})^T,$$

then

$$\vec{p} = M_1 \times \vec{r}. \quad (1)$$

Here  $M_1$  is a 7 by 7 matrix determined by the above seven equations derived under Model (1). It is straightforward to verify that the probability of catching any individual in each sample is fixed, i.e.,  $p_{1++} = r_{1++} = p_1$ ,  $p_{+1+} = r_{+1+} = p_2$ ,  $p_{++1} = r_{++1} = p_3$ . This must be the case because the sample capture probabilities do not depend on how the matching mechanism operates.

We can easily generalize this formulation to handle the  $k$ -sample case; however, the algebra involved is quite messy for large  $k$ . We can simplify this model by requiring that the transitions can go downwards by at most one level, thus yielding Model (2):

**Model (2).** In addition to the homogeneity and closure assumptions in §1, we assume that: (i) there are no erroneous matches in the matching process; (ii) a transitions can only go downwards by at most one level; (iii) any individual will stay at his original state with probability  $\theta$ , and transition to any of a possible set of individuals with probability  $(1 - \theta)/(m' - 1)$ , where  $m'$  is the number of sets of individuals to which transitions are possible and allowed.

We first consider the three-sample case. A (1,1,1) individual can decompose into three individuals, i.e.,  $(1,1,1) \mapsto \{(1,0,0), (0,1,0), (0,0,1)\}$  (we use " $\mapsto$ " to denote for decomposition), if three presumed matches are not made. Assumption (ii) of Model (2) assumes that this triple error has negligible probability when compared with the transition in which only one of the matches is not made so that  $(1,1,1) \mapsto \{(1,1,0), (0,0,1)\}$ , or  $(1,1,1) \mapsto \{(1,0,1), (0,1,0)\}$ , or  $(1,1,1) \mapsto \{(1,1,0), (0,0,1)\}$ .

For three sample case, the parametric model for expressing  $\{p_{ijk}\}$  in terms of  $\{r_{ijk}\}$  is:

$$p_{111} = \theta r_{111},$$

$$p_{112} = \frac{1 - \theta}{3} r_{111} + \theta r_{112},$$

$$p_{121} = \frac{1 - \theta}{3} r_{111} + \theta r_{121},$$

$$p_{211} = \frac{1 - \theta}{3} r_{111} + \theta r_{211},$$

$$p_{122} = \frac{1-\theta}{3} r_{111} + (1-\theta)r_{112} + (1-\theta)r_{121} + r_{122},$$

$$p_{212} = \frac{1-\theta}{3} r_{111} + (1-\theta)r_{112} + (1-\theta)r_{211} + r_{212},$$

$$p_{221} = \frac{1-\theta}{3} r_{111} + (1-\theta)r_{211} + (1-\theta)r_{121} + r_{221}.$$

Then

$$\vec{p} = M_2 \times \vec{r}, \quad (2)$$

where  $M_2$  is a 7 by 7 matrix determined by the above seven equations derived under Model (2). Again, the capture probabilities are unchanged, *i.e.*,  $p_{1++} = r_{1++} = p_1$ ,  $p_{+1+} = r_{+1+} = p_2$ ,  $p_{++1} = r_{++1} = p_3$ .

For the  $k$ -sample problem, let  $p_{\bar{1}}$  be the probability of being captured in all samples, *i.e.*,  $p_{\bar{1}} = p_{111\dots 1}$ , and let  $p_{\bar{1}, \bar{2}(h_1, h_2)}$  be the cell probability corresponding to absence in the  $h_1$ -th, and  $h_2$ -th sample and presence in the others, *etc.* Under Model (2), we have  $p_{\bar{1}} = \theta r_{\bar{1}}$ . For  $i \leq k-2$ , the probability of being missed by the  $h_1$ -th,  $h_2$ -th,  $\dots$ , and  $h_i$ -th sample and captured by the others is

$$p_{\bar{1}, \bar{2}(h_1, h_2, \dots, h_i)} = \theta r_{\bar{1}, \bar{2}(h_1, h_2, \dots, h_i)} + \frac{1-\theta}{k-i+1} \sum_{j=1}^i r_{\bar{1}, \bar{2}(\{h_1, h_2, \dots, h_i\} \setminus h_j)}.$$

For  $i = k-1$ , the individual is included in only one sample. For example, the probability of being captured only by the first sample is

$$p_{1, \bar{2}} = r_{1, \bar{2}} + (1-\theta) \sum_{h \neq 1} r_{1, 1(h), \bar{2}} + \frac{(1-\theta)}{3} \sum_{h_1, h_2 \geq 2} r_{1, 1(h_1, h_2), \bar{2}} + \sum_{j=3}^{k-1} \sum_{h_1, h_2, \dots, h_j \geq 2} \frac{(1-\theta)}{(j+1)} r_{1, 1(h_1, h_2, \dots, h_j), \bar{2}},$$

where  $r_{1, 1(h_1, h_2, \dots, h_j), \bar{2}}$  is the cell probability in the original table which corresponds to presence in the first,  $h_1$ -th,  $h_2$ -th,  $\dots$ ,  $h_j$ -th sample and absence in the others. By symmetry, we can write down the expression for  $p_{1(h), \bar{2}}$ , the probability of being observed in the  $h$ -th sample only and missed in all others.

We can refine Model (2) by assuming unequal matching rates. For example, we consider two decompositions:  $(1, 1, 1) \mapsto \{(1, 1, 0), (0, 0, 1)\}$  and  $(1, 1, 0) \mapsto \{(0, 1, 0), (1, 0, 0)\}$ .

It is common for both cases that one presumed match is not made. They differ in that one has two sources of information for that match while the other has only one. It is reasonable to assume different matching error probabilities for the two cases instead of a common one as proposed in Model (2). This leads to:

**Model (3).** In addition to (i) and (iii) in Model (2), we assume

$$(1, 1, 1) \mapsto \begin{cases} (1, 1, 1) & \text{with probability } \alpha_1 \\ \{(1, 1, 0), (0, 0, 1)\} & \text{with probability } (1-\alpha_1)/3 \\ \{(0, 1, 1), (1, 0, 0)\} & \text{with probability } (1-\alpha_1)/3 \\ \{(1, 0, 1), (0, 1, 0)\} & \text{with probability } (1-\alpha_1)/3 \end{cases}$$

$$(1, 1, 0) \mapsto \begin{cases} (1, 1, 0) & \text{with probability } \alpha_2 \\ \{(0, 1, 0), (1, 0, 0)\} & \text{with probability } 1-\alpha_2 \end{cases}$$

$$(1, 0, 1) \mapsto \begin{cases} (1, 0, 1) & \text{with probability } \alpha_2 \\ \{(1, 0, 0), (0, 0, 1)\} & \text{with probability } 1-\alpha_2 \end{cases}$$

$$(0, 1, 1) \mapsto \begin{cases} (0, 1, 1) & \text{with probability } \alpha_2 \\ \{(0, 1, 0), (0, 0, 1)\} & \text{with probability } 1-\alpha_2 \end{cases}$$

and  $(1, 0, 0)$ ,  $(0, 1, 0)$ ,  $(0, 0, 1)$  stay the same with probability one.

Under this model, we can express the cell probability  $\{p_{ijk}\}$  in Table 2 in terms of  $\alpha_1$ ,  $\alpha_2$  and the cell probabilities of Table 1,  $\{r_{ijk}\}$ . To do this, we need to consider all possible transitions that produce an individual that falls into the  $(i, j, k)$  cell in Table 2. For example, we consider an observed  $(1, 0, 0)$  individual. This person falls into cell  $(1, 2, 2)$  of Table 2. Let  $F$  be the event that an observed individual has a  $(1, 0, 0)$  status. Let  $E_{ijk}$  be the event that an individual falls into  $(i, j, k)$  cell in Table 1. Then

$$F = \bigcup_{\{i, j, k\}} (E_{ijk} \cap F).$$

According to Model (3), there are only four possible transitions as follows that can make  $F$  happen:

$$(1, 1, 1) \mapsto \{(1, 0, 0), (0, 1, 1)\},$$

$$(1, 1, 0) \mapsto \{(1, 0, 0), (0, 1, 0)\},$$

$$(1, 0, 1) \mapsto \{(1, 0, 0), (0, 0, 1)\},$$

$$(1, 0, 0) \mapsto \{(1, 0, 0)\}.$$



Therefore

$F =$

$$(E_{111} \cap F) \cup (E_{112} \cap F) \cup (E_{121} \cap F) \cup (E_{122} \cap F).$$

By the definitions of cell probabilities of the two tables,  $p(F) = p_{122}$ , and  $p(E_{ijk}) = r_{ijk}$ . By the assumptions in Model (3),  $p(F | E_{111}) = (1 - \alpha_1)/3$ ,  $p(F | E_{112}) = p(F | E_{121}) = \alpha_2$ , and  $p(F | E_{122}) = 1$ .

Since  $E_{111} \cap F$ ,  $E_{112} \cap F$ ,  $E_{121} \cap F$  and  $E_{122} \cap F$  are four mutually exclusive possibilities that  $F$  can happen, thus

$$\begin{aligned} p_{122} &= p(E_{111} \cap F) + p(E_{112} \cap F) \\ &\quad + p(E_{121} \cap F) + p(E_{122} \cap F) \\ &= p(F | E_{111}) \cdot p(E_{111}) + p(F | E_{112}) \cdot p(E_{112}) \\ &\quad + p(F | E_{121}) \cdot p(E_{121}) + p(F | E_{122}) \cdot p(E_{122}) \\ &= \frac{1 - \alpha_1}{3} r_{111} + (1 - \alpha_2) r_{112} + (1 - \alpha_2) r_{121} + r_{122}. \end{aligned}$$

In the same manner, we can derive the expressions of other cell probabilities of Table 2 to get

$$p_{111} = \alpha_1 r_{111},$$

$$p_{112} = \frac{1 - \alpha_1}{3} r_{111} + \alpha_2 r_{112},$$

$$p_{121} = \frac{1 - \alpha_1}{3} r_{111} + \alpha_2 r_{121},$$

$$p_{211} = \frac{1 - \alpha_1}{3} r_{111} + \alpha_2 r_{211},$$

$$p_{122} = \frac{1 - \alpha_1}{3} r_{111} + (1 - \alpha_2) r_{112} + (1 - \alpha_2) r_{121} + r_{122},$$

$$p_{212} = \frac{1 - \alpha_1}{3} r_{111} + (1 - \alpha_2) r_{112} + (1 - \alpha_2) r_{211} + r_{212},$$

$$p_{221} = \frac{1 - \alpha_1}{3} r_{111} + (1 - \alpha_2) r_{211} + (1 - \alpha_2) r_{121} + r_{221}.$$

Then

$$\vec{p} = M_3 \times \vec{r}, \quad (3)$$

where  $M_3$  is a 7 by 7 matrix determined by the above seven equations derived under Model (3).

For  $\alpha_1 = \alpha_2 = \theta$ , we get the same formulation as under Model (2). For the special case with  $\alpha_1 = \alpha_2 = 1$ ,  $p_{ijk} = r_{ijk}$ , reducing to the traditional problem. Again, the capture probabilities remain the same, i.e.,  $p_{1++} = r_{1++}$ ,  $p_{+1+} = r_{+1+}$ ,  $p_{++1} = r_{++1}$ .

## 4. ESTIMATING THE SIZE OF THE POPULATION

### 4.1 Log-linear Model Formulation

For purposes of exposition, we confine our attention to the three-sample census case, although extensions to the  $k$ -sample census for  $k > 3$  are straightforward. As before, let  $l_{ijk}$  and  $m_{ijk}$  be expected cell counts for Table 1 and Table 2 respectively. The relationship between the cell probabilities and the expected cell counts is  $l_{ijk} = r_{ijk}N$ , and  $m_{ijk} = p_{ijk}N$ . Let

$$\vec{m} = (m_{111}, m_{112}, m_{121}, m_{211}, m_{122}, m_{212}, m_{221})^T,$$

and

$$\vec{l} = (l_{111}, l_{112}, l_{121}, l_{211}, l_{122}, l_{212}, l_{221})^T.$$

Since for each of the models we have proposed in the last section, there is a matrix  $M$  with entries depending on the matching probability parameters in the chosen model such that  $\vec{p} = M \times \vec{r}$ , multiplying through by  $N$  gives

$$\vec{m} = M \times \vec{l}. \quad (4)$$

For any log-linear model specified for Table 1, it is straightforward to obtain the parameterization for  $m_{ijk}$ . For example, for any of the models suggested in Fienberg (1972), we can write the expected counts in terms of functions of  $u$ -term parameters:

$$l_{ijk} = g_{ijk}(u, u_1(i), u_2(j), u_3(k), u_{12}(ij), u_{13}(ik), u_{23}(jk)), \quad (5)$$

and then obtain the parameterization of  $\{m_{ijk}, (ijk) \neq (222)\}$  from (4).

### 4.2 Estimating the Size of the Population

We now consider the matching rates in our various models as known. To obtain the estimate of the population size, we proceed as follows. First, following Sanathanan (1972), we compute the maximum likelihood estimates of  $u$ -term parameters from  $l_c$ , the conditional likelihood associated with Table 2 given  $n$ ,

$$l_c = n! \prod_{\{(ijk) \neq (222)\}} \frac{(q_{ijk})^{x_{ijk}}}{x_{ijk}!},$$

where  $n = \sum_{\{(ijk) \neq (222)\}} x_{ijk}$ , and  $q_{ijk} = m_{ijk}/n$ . Sanathanan (1972) shows that, under suitable regularity conditions, the conditional maximum likelihood estimates and the unconditional ones are both consistent and have the same asymptotic normal distribution. If we remove redundant  $u$ -term parameters using the constraints associated with the specified log-linear model for Table 1, then the problem is to find the maximum of  $l_c$  subject to the following single constraint:

$$\sum_{\{(ijk) \neq (222)\}} m_{ijk} = n.$$

Numerically, this is a nonlinearly constrained optimization problem. Rao (1957) studied regularity conditions under which there exist unique maximum likelihood estimates of the parameters in a multinomial distribution. His conditions are satisfied by the parameterization of  $\{q_{ijk}\}$ . Once the conditional maximum likelihood estimates of the  $u$ -term parameters are obtained, we use the loglinear model specified for Table 1 to compute the conditional maximum likelihood estimates of  $\{l_{ijk}\}$ , the expected cell counts of Table 1 including the expected count of the missing cell. Then our estimate of  $N$  is

$$\hat{N} = \sum_{\{ijk\}} \hat{l}_{ijk}.$$

In the case of no matching errors, with  $\alpha_1 = \alpha_2 = 1$  in Model (3),  $m_{ijk} = l_{ijk}$ . Thus

$$\hat{N} = n + \hat{m}_{222},$$

*i.e.*, we get back to the estimation method for the traditional multiple recapture census problem developed by Fienberg (1972) when the log-linear models in Fienberg (1972) are considered.

As we have discussed earlier, a log-linear model is specified for Table 1 and the observations are viewed as falling into Table 2, whose parametric model of the expected cell counts is specified by the log-linear model and a chosen model for matching errors. To assess the appropriateness of a log-linear model specified for Table 1, we can apply the usual Pearson and likelihood ratio goodness-of-fit tests,  $X^2$  and  $G^2$ , discussed in Fienberg (1972), to Table 2. Each statistic has an asymptotic  $\chi^2$  distribution under the null hypothesis that the model fits, with degrees of freedom equal to  $2^k - 1 - (\text{number of independent parameters in the model})$ .

## 5. ANALYSIS OF 1988 ST. LOUIS DRESS REHEARSAL CENSUS DATA

Dual System Estimation (DSE), based on the standard two-sample census, has been employed by U.S. Bureau of Census for census coverage evaluation since 1950. In 1988,

the Census Bureau conducted a Dress Rehearsal Census for the 1990 decennial census at three sites: St. Louis, Missouri; Columbia, Missouri; and western Washington State. Zaslavsky and Wolfgang (1993) present data for a population subgroup from the Post Enumeration Survey (PES) in the dress rehearsal census in St. Louis which focuses on urban Black male adults who are believed to be underestimated by dual system methods. The resulting data consists of three sources: the  $C$ -sample is the census itself; the  $P$ -sample was compiled from the PES; a third source of information was the Administrative List Supplement (ALS), compiled from pre-census administrative records of state and federal government agencies, encompassing Employment Security, driver's license, Internal Revenue Service, Selective Service, and Veteran's Administrative records. The  $C$ -sample and  $P$ -sample provide data for the implementation of the usual DSE or capture recapture approach. The ALS data can be combined with the Census and the  $P$ -sample for analysis from a three-sample perspective, though it was originally intended to improve the coverage of the  $P$ -sample. In Table 3, we present three-sample data for PES sampling stratum 11 in St. Louis obtained by collapsing the original data in Table 1 of Zaslavsky and Wolfgang (1993) over four poststrata defined by owners/renters  $\times$  age 20-29, 30-44.

Table 3

Three-Sample Data for Stratum 11, St. Louis

ALS	Census			
	Present $P$ -sample		Absent $P$ -sample	
	Present	Absent	Present	Absent
	Present	Absent	Present	Absent
Present	300	51	53	180
Absent	187	166	76	—

Such triple-system data can be analyzed with the matching error Model (2) and data from a separate Matching Error Study (MES, or rematch study) associated with the same sampling poststratum. The MES is one of the operations conducted by the Census Bureau to evaluate the PES, and typically operates for a sample of cases, using more extensive procedures, highly qualified personnel and reinterviews to obtain estimates of the bias associated with the previous matching process. In the discussion of the Matching Error Study done in a 1986 test census in Los Angeles, Hogan and Wolter (1988) state that "The rematch was done independently of the original match, and the discrepancies between the match and the rematch results are adjudicated. Because of this intensive approach to the rematch, we believe the rematch results represent true match status, while differences between the match and rematch results represent the bias in the original match results."



Table 4

St. Louis Rematch Study: *P*-sample  
Source: Mulry, Dajani and Biemer (1989)

Original Match Classification	Rematch Classification			Total
	Matched	Not Matched	Un-resolved	
Matched	2,667	7	8	2,682
Not matched	9	427	30	466
Unresolved	0	7	20	27
Total	2,676	441	58	3,175

The data from the MES thus provides a basis for estimating error rates in the original matching process. Mulry, Dajani and Biemer (1989) report the MES operation for the 1988 Dress Rehearsal and rematch data for all three test sites, and in Table 4, we reproduce those data relevant for our purposes.

Let  $\alpha$  be the matching rate between the *C*-sample and the *P*-sample, and  $\gamma = 1 - \alpha$  be the nonmatch error rate. We assume no errors in the rematch. Then from the data in Table 4, we can estimate  $\alpha$  by  $\hat{\alpha} = 2667/(2667 + 9) = 99.6637\%$ , and  $\gamma$  by  $\hat{\gamma} = 1 - \hat{\alpha} = .3363\%$ . The parameter  $\theta$  is a three-sample matching rate for the *C*-sample, *P*-sample and the ALS. It takes two matches, say, one between the *C*-sample and the *P*-sample, and the other one between the *P*-sample and the ALS, in order to reach a correct (1,1,1) three-sample classification. In the absence of evaluation of the match between the census and the ALS, we assume that these two matches are independent of each other and that the matching rate for the *P*-sample and ALS is the same for the *C*-sample and the *P*-sample. Thus we can use  $\theta = \alpha^2$ , and  $\hat{\theta} = \hat{\alpha}^2 = 99.3285\%$ . Based on other qualitative information, this seems to be unreasonably high match rate, and the match error rate for the census and the ALS is probably higher than the match error rate between the census and the *P*-sample. In the absence of better quantitative information, however, we proceed to use it in the calculations that follow.

Table 5

Estimates Under Various Models

Log-linear Model	Usual MLE		MLE Using Matching Error Model (2)	
	$\hat{N}$ (S.E.)	Fit (d.f.)	$\hat{N}$ (S.E.)	Fit (d.f.)
[C] [P] [A]	1091.48 (11.24)	248.31 (3)	1083.58 (10.93)	244.56 (3)
[CP] [A]	1204.14 (23.31)	90.60 (2)	1194.73 (22.86)	87.30 (2)
[PA] [C]	1108.34 (13.77)	247.93 (2)	1100.03 (13.40)	244.53 (2)
[CA] [P]	1068.87 (10.47)	230.66 (2)	1061.09 (10.10)	226.42 (2)
[CP] [CA]	1271.11 (52.55)	87.16 (1)	1256.77 (50.97)	84.37 (1)
[CP] [PA]	1598.88 (106.26)	17.55 (1)	1585.03 (104.93)	15.88 (1)
[CA] [PA]	1080.47 (13.38)	230.43 (1)	1072.19 (12.88)	226.44 (1)
[CP] [CA] [PA]	2360.82 (363.25)	— (0)	2309.55 (352.36)	— (0)

Table 5 gives the estimates of the population size for various log-linear models with estimates of standard errors and goodness-of-fit statistics. Standard errors are computed with the delta method as discussed in Fienberg (1972). The assumption of independence between the census and the *P*-sample has been questioned for the use of the DSE. The dual system method has limited capacity to test this assumption and to adjust for potential dependency, while both can be handled through log-linear models for three or more samples. There are four models listed in Table 5 that assume independence between the census and the *P*-sample: the independence model [C] [P] [A], [PA] [C], [CA] [P], and [CA] [PA]. All of them fit the data poorly. The three models with the interaction term for the census and the *P*-sample, [CP] [A], [CP] [CA], and [CP] [PA] fit the data much better. With the addition of an interaction term linking the census and the ALS, model [CP] [CA] fits only slightly better than [CP] [A], indicating that the census and the *P*-sample are together nearly independent from the ALS. The model [CP] [PA] fits the data the best, suggesting that the usual independence assumption for the DSE is invalid and that there is dependence between the *P*-sample and the ALS. For all seven non-saturated log-linear models, we obtain better fits under matching error Model (2), though only slightly so, due to the high match rate for the data from the 1988 U.S. Census Dress Rehearsal. For the [CP] [PA] model, there is a .8738% difference in the estimate of *N* associated with the nonmatch rate of .3363%. If the nonmatch rate had been 10%, *i.e.*, a 90% match rate, and assuming that the difference in the estimate of *N* is approximately linear in the nonmatch rate, there would have been a 26% difference between the usual maximum likelihood estimate of *N* and our estimate.

Table 6

Dual-System Data for Stratum 11, St. Louis

<i>P</i> -sample	Census		Total
	Present	Absent	
Present	487	129	616
Absent	217	—	
Total	704		

Table 6 presents the usual dual system data for stratum 11, St. Louis. The number of people in both the census and the *P*-sample is  $y_{11} = 300$ , the number of those in the census only is  $y_{12} = 217$ , and number in the *P*-sample only is  $y_{21} = 129$ . The total census count is  $y_{1+} = y_{11} + y_{12} = 704$ , the total *P*-sample count is  $y_{+1} = y_{11} + y_{21} = 616$ , the dual system estimate is  $DSE = y_{1+}y_{+1}/y_{11} = 893$  (p. 232, Bishop, Fienberg and Holland 1975), and the estimated variance of  $DSE$  is  $Var(DSE) = y_{1+}y_{+1}y_{12}y_{21}/y_{11}^3 = 105.4$  (p. 233, Bishop *et al.* 1975). The standard error is  $SE(DSE) = 10.27$ .

The census undercount for the population estimate  $\widehat{DSE}$  is  $(\widehat{DSE} - y_{1+}) / \widehat{DSE} \times 100\% = 21.17\%$ . For our best fitting model, the census undercount is  $(\hat{N} - y_{1+}) / \hat{N} = 55.97\%$  for the estimate  $\hat{N} = 1599$  assuming no matching error and 55.58% for  $\hat{N} = 1585$  from matching error Model (2). Thus there is a 55.97% – 55.58% = 0.39% upward bias by ignoring matching errors. This is quite close to the figure of 0.37% computed in Ding and Fienberg (1994) for the 1986 Los Angeles test census data using a two-sample match rate of 99.4734%, as compared to 99.6637% here for the St. Louis data. Our estimates show that the urban Black male adults targeted in the St. Louis Dress Rehearsal were heavily undercounted by the census, and that the undercount is severely underestimated by the usual dual-system or capture-recapture estimator of the population size. A third and qualitatively different sample might work well for this demographic group.

The homogeneity of the capture probabilities is one of the assumptions in the standard approach to the estimation of the size of a closed population. Darroch *et al.* (1993) developed a quasi-symmetry model and a partial quasi-symmetry model to allow for varying catchability of individuals. The quasi-symmetry model assumes that the pattern of heterogeneity is the same for all three samples, the partial quasi-symmetry model assumes that the pattern of heterogeneity is the same for two samples but different for the third sample. This is a sensible model given that the third sample is qualitatively quite different from the census and the PES and this model is equivalent to a combination of dependence and heterogeneity. For the multinomial cell probabilities including the missing cell,  $R = (r_{111}, r_{112}, \dots, r_{222})$ , both are log-linear models of the form  $\log R = A\beta$  with an appropriately chosen design matrix  $A$  and a vector of parameters  $\beta$ . The design matrices for both models are given in Darroch *et al.* (1993).

**Table 7**  
Heterogeneous Catchability Models

Log-Linear Model	MLE from Darroch <i>et al.</i> (1993)		MLE Using Matching Error Model (2)	
	$\hat{N}$ (S.E.)	Fit (d.f.)	$\hat{N}$ (S.E.)	Fit (d.f.)
Full quasi-symmetry	1923.63 (216.84)	133.54 (2)	1906.61 (213.47)	133.50 (2)
Partial quasi-symmetry	2576.54 (413.28)	11.70 (1)	2557.08 (409.39)	11.72 (1)

Our proposed method can readily incorporate heterogeneous catchability to estimate the population size by assuming a heterogeneity model for Table 1 and then adopting the conditional likelihood estimation (Sanathanan 1972). Table 7 presents estimates from fitting the quasi-symmetry model and the partial quasi-symmetry model for

the data from stratum 11. Again, the effect of the matching errors in this analysis is not substantial due to the high matching rate. The partial quasi-symmetry model fits much better than the quasi-symmetry model, indicating there seems to be plausible heterogeneity and the pattern of heterogeneity seems different in the ALS. The lack of fit of the independence model might also be explained in part by the dependence among the samples (in particular between the census and the  $P$ -sample) and in part by heterogeneous catchability.

The partial quasi-symmetry model incorporates the [CP] dependence and thus is an alternative to the model [CP] [PA] in Table 5. The two models yield similar fits to the data, but they give dramatically different estimates of  $N$ , with the model incorporating heterogeneity having a much larger estimate accompanied by a much larger estimated standard error. This suggests that there is a considerable instability associated with heterogeneity parameters and, although the two models are not nested and thus not directly comparable, it seems reasonable to opt for the smaller and more stable estimate which does not incorporate heterogeneity.

Darroch *et al.* (1993) considered four substrata for stratum 11 in their analysis. The two cross-classification variables for the four substrata O2, R2, O3 and R3 are whether residents owned or rented homes and whether they were age 20-29 or 30-44. The data for the four substrata are given in Table 8 where 1 corresponds to presence in a sample and 0 is for absence. We have reanalyzed them for comparison. Table 9 and Table 10 give estimates for both heterogeneity models. As pointed out earlier, the high match rate yields similar estimates and fits for models incorporating matching errors. The partial quasi-symmetry model shows significant improvement in fits over the full quasi-symmetry model with the best fits obtained for R2 and R3. If we add the estimates of  $N$  across the four substrata, the total for the matching error version of partial quasi-symmetry is  $\hat{N} = 2980.8$ , more than 16% larger than the estimate from the collapsed model in Table 7. Of course, the standard error of the estimate has increased by a similar magnitude.

**Table 8**  
Three-Sample Data for Four Substrata of Stratum 11  
Source: Table 2, Darroch *et al.* (1993)

Sample			Substratum			
$C$	$P$	$A$	O2	R2	O3	R3
0	0	1	59	43	35	43
0	1	0	8	34	10	24
0	1	1	19	11	10	13
1	0	0	31	41	62	32
1	0	1	19	12	13	7
1	1	0	13	69	36	69
1	1	1	79	58	91	72



**Table 9**  
Estimates for Full Quasi-Symmetry

Sub-stratum	MLE from Darroch <i>et al.</i> (1993)		MLE Using Matching Error Model (2)	
	$\hat{N}$ (S.E.)	Fit (d.f.)	$\hat{N}$ (S.E.)	Fit (d.f.)
O2	780.83 (294.81)	11.70 (2)	777.98 (293.99)	11.69 (2)
R2	394.34 (56.45)	41.09 (2)	391.14 (55.29)	41.02 (2)
O3	765.45 (254.57)	25.99 (2)	759.97 (252.44)	25.98 (2)
R3	361.83 (47.33)	59.31 (2)	358.71 (46.20)	59.22 (2)

**Table 10**  
Estimates for Partial Quasi-Symmetry

Sub-stratum	MLE from Darroch <i>et al.</i> (1993)		MLE Using Matching Error Model (2)	
	$\hat{N}$ (S.E.)	Fit (d.f.)	$\hat{N}$ (S.E.)	Fit (d.f.)
O2	605.66 (212.63)	7.51 (1)	601.44 (210.93)	7.52 (1)
R2	652.34 (205.12)	0.04 (1)	646.59 (202.58)	0.04 (1)
O3	1124.00 (473.26)	8.27 (1)	1126.90 (476.54)	8.22 (1)
R3	611.78 (200.82)	2.92 (1)	605.91 (198.26)	2.92 (1)

## 6. SUMMARY

In this paper, we have presented models for matching errors and models for the estimation of the population total and census undercount in a multiple sample census. We have illustrated our methods by reanalyzing census coverage data from the 1988 St. Louis Dress Rehearsal census. Two sources of information are considered in our analysis, the data from a Matching Error Study (MES), and triple-system data with every individual cross-classified according to presence or absence in each of three samples: the census, a post enumeration survey (*P*-sample) and an administrative list supplement. We imbed the standard log-linear model formulation of Fienberg (1972) into our estimation procedure to account for statistical dependency together with matching errors and to allow for formal goodness-of-fit test of various models. Our method applies to any model of a log-linear form and we have illustrated how heterogeneity models can be incorporated into our approach to allow for both matching errors and heterogeneous catchability.

Our matching error models assume that false matches are negligible. Sensitivity analysis in Ding (1990) shows that when both the false nonmatch rate and the false match rate are the same order of magnitude, the matching bias is dominated by the false nonmatch rate (see also Fay, Passel, Robinson and Cowan 1988, p. 53). This is because the capture probabilities in the census and the post enumeration

survey are high, and thus a comparable change in both the false nonmatch and false match rates has substantially more impact on false nonmatches than false matches. For the 1986 Los Angeles test census data, the estimates of false nonmatch rate and false match rate computed in Ding and Fienberg (1994) are about 0.5% and 0.8%, respectively. Based on these empirical findings, we have some reason to believe that, at least in the census application described here, our models for false nonmatch errors are reasonable approximations to reality.

We have analyzed the St. Louis triple-system data with an estimate of the matching rate taken from the MES. Matching rates may not be homogeneous over different population strata, and we suggest that the MES data associated with the same sampling stratum be used. We have developed formulation in §3 for the *k*-sample census, and our approach can be readily applied to a *k*-sample census with  $k \geq 4$ .

## REFERENCES

- BISHOP, Y.M.M., FIENBERG, S.E., and HOLLAND, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: M.I.T. Press.
- CHEN, T.T. (1979). Log-linear models for categorical data with misclassification and double sampling. *Journal of American Statistical Association*, 74, 481-488.
- CORMACK, R.M. (1968). The statistics of capture-recapture methods. *Oceanography and Marine Biology, Annals Review*, 6, 455-506.
- DARROCH, J.N. (1958). The multiple-recapture census, I: estimation of a closed population. *Biometrika*, 45, 343-359.
- DARROCH, J.N., FIENBERG, S.E., GLONEK, G.F.V., and JUNKER, B.W. (1993). A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. *Journal of American Statistical Association*, 88, 1137-1148.
- DING, Y. (1990). Capture-recapture census with uncertain matching. Ph.D. dissertation, Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania.
- DING, Y., and FIENBERG, S.E. (1994). Dual system estimation of census undercount in the presence of matching error. *Survey Methodology*, 20, 149-158.
- FAY, R.E., PASSEL, J.S., ROBINSON, J.G., and COWAN, C.D. (1988). The coverage of population in the 1980 census. Bureau of the Census, U.S. Department of Commerce.
- FIENBERG, S. E. (1972). The multiple recapture census for closed populations and incomplete  $2^k$  contingency tables. *Biometrika*, 59, 591-603.
- HOGAN, H., and WOLTER, K. (1988). Measuring accuracy in a Post-Enumeration Survey. *Survey Methodology*, 14, 99-116.
- MULRY, M.H., DAJANI, A., and BIEMER, P. (1989). The Matching Error Study for the 1988 Dress Rehearsal. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 704-709.

- RAO, C.R. (1957). Maximum likelihood estimation for the multinomial distribution. *Sankhyā*, 18, 139-148.
- SANATHANAN, L. (1972). Estimating the size of a multinomial population. *Annals of Mathematical Statistics*, 43, 142-152.
- SEBER, G.A.F. (1982). *The Estimation of Animal Abundance and Related Parameters*. New York: MacMillan.
- ZASLAVSKY, A.M., and WOLFGANG, G.S. (1993). Triple System Modeling of Census, Post-Enumeration Survey and Administrative List Data. *Journal of Business and Economic Statistics*, 11, 279-288.



# Applying the Lavallée and Hidiroglou Method to Obtain Stratification Boundaries for the Census Bureau's Annual Capital Expenditures Survey

JOHN G. SLANTA and THOMAS R. KRENZKE<sup>1</sup>

## ABSTRACT

The Lavallée-Hidiroglou (L-H) method of finding stratification boundaries has been used in the Census Bureau's Annual Capital Expenditures Survey (ACES) to stratify part of its universe in the pilot study and the subsequent preliminary survey. This iterative method minimizes the sample size while fixing the desired reliability level by constructing appropriate boundary points. However, we encountered two problems in our application. One problem was that different starting boundaries resulted in different ending boundaries. The other problem was that the convergence to locally-optimal boundaries was slow, *i.e.*, the number of iterations was large and convergence was not guaranteed. This paper addresses our difficulties with the L-H method and shows how they were resolved so that this procedure would work well for the ACES. In particular, we describe how contour plots were constructed and used to help illustrate how insignificant these problems were once the L-H method was applied. This paper describes revisions made to the L-H method; revisions that made it a practical method of finding stratification boundaries for ACES.

**KEY WORDS:** Convergence; Contour plots; Economic surveys.

## 1. INTRODUCTION

The primary objectives of the sample design of the Census Bureau's Annual Capital Expenditures Survey (ACES) are to meet desired reliability levels using operationally-feasible methodology and to stay within budget limitations. To achieve these goals, we implemented a stratified simple random sample design using a modified version of Lavallée and Hidiroglou's (L-H) (1988) approach of finding stratum bounds. This stratification method for skewed populations obtains optimal boundary points by minimizing the total sample size given a desired coefficient of variation (c.v.). Survey managers associated with a single-purpose survey having access to a single stratifier can benefit from its operational ease and cost reductions.

We considered several papers that documented other methods for finding size stratum boundaries. Hess, Sethi, and Balakrishnan (1966) compared several stratifying techniques. The popular Dalenius and Hodges method (Cochran 1977, p. 129) was considered easy to implement in our case but was initially ruled out because it was not designed with certainty strata in mind. Sethi's method (1963) of using standard distributions was not used because we thought it would be cumbersome to identify the distribution and sub-optimal to use standard distributions for each of the 80 ACES industries. Eckman's rule (1959) of equalizing the product of stratum weights and stratum range seemed to require rather ominous calculations.

The L-H method was the most appealing to our application. Designed specifically for skewed populations, which is often the case for economic surveys, it creates a boundary that defines the take-all stratum, and the optimal boundary point(s) for the take-some strata. It sometimes will create additional take-all strata if through Neyman Allocation, the stratum sample size is greater than or equal to the stratum size.

The L-H method goes through an iterative algorithm beginning with computing or arbitrarily setting the initial stratum boundaries. Then, stratum statistics are computed such as, the stratum size, mean, and the variance. These parameters are entered into boundary formulas that were derived from minimizing the sample size subject to a desired cv. If the new boundaries do not converge then the stratum statistics are calculated for the newly defined size strata. The cycle continues until the boundaries converge.

Schneeberger (1979) discussed the problem of finding optimal stratification boundaries. Schneeberger shows in the paper that when expressing this problem as a non-linear program, when solved by a gradient method, the solution may be relative or global minima, maxima, or saddle points of the variance of the sample mean. Detlefsen and Veum (1991) document this as a shortcoming of the L-H method when testing its application for the Census Bureau's Monthly Retail Trade Survey. In the L-H method, they found that many times the resulting boundaries differed substantially from where the initial boundaries were set,

<sup>1</sup> John G. Slanta, Manufacturing and Construction Division; Thomas R. Krenzke, Decennial Statistical Studies Division, U.S. Bureau of the Census, Washington, D.C. 20233, U.S.A.

so the minimum sample size attained was a local minimum. Geometrically, the sample size as a function of two strata boundaries, appears like a landscape with one or more bowl-shaped valleys. The L-H method begins in a region and descends until it reaches the lowest point. If more than one minimum exists, it will not continue to search for the global minimum. Therefore, one objective is to have initial boundaries that are in the neighborhood of the global minimum. Using starting boundaries resulting from a technique such as the Dalenius and Hodges method may help satisfy this desire.

Detlefsen and Veum (1991) also noted instances of slow or non-convergence. However, they also noted that convergence occurred faster when the number of strata was reduced and when starting boundaries were the same as the previous survey's sample selection boundaries. In order to defend ourselves against infinite loops in the computer program or a large number of iterations, we decided on doing two things. First, we implemented a sample design in which the L-H method would create sets of only three size strata. Second, we decided to implement stopping rules so that when the convergence rate appeared to slow down, the program stopped processing.

In this work, we give background information on the ACES and briefly describe the way the L-H method was applied. We show how contour plots and three-dimensional plots gave us justification for using the L-H method to get the final boundaries. We show how the contour plots address the convergence problem by showing how constraints can be setup to be met after each iteration. This would protect us against slow or non-convergence under the assumption that the marginal gain achieved is not worth the extra effort.

## 2. ACES BACKGROUND

The 1992 ACES was designed by the Census Bureau to be a large-scale operational test of the sampling, processing, programming, data entry, editing, and estimation procedures which extended beyond a 1991 pilot study, to prepare for the 1993 full-scale survey. Capital expenditure estimates for domestic activities were published at conglomerated industry levels from the 1992 survey. In addition, the 1991 and 1992 preliminary surveys provided valuable capital expenditure data that will be used in future sample design enhancements.

The sampling unit for the ACES was the company which may be comprised of several establishments. The sampled population included all active companies with five or more employees from all major industry sectors except Government. These sectors include mining, construction, manufacturing, transportation, wholesale and retail trade, finance, services, and a portion of the agriculture sector that includes agricultural services, forestry, fishing,

hunting, and trapping. Only companies with domestic activity were included in the sampling frame. The Research and Methodology Staff of the Census Bureau's Industry Division constructed the sampling frame, selected the sample, and generated estimates.

The ACES sampling frame was constructed from the Census Bureau's Standard Statistical Establishment List (SSEL) in November 1992 using final 1991 data for single unit (SU) establishments and 1990 data for establishments associated with multiunit (MU) firms. Major exclusions from the frame were public administration, U.S. Postal Service, international establishments, establishments in Puerto Rico, Guam, Virgin Islands, and the Mariana Islands. EI Submasters which are SU records on the SSEL that are associated with MU establishments, establishments associated with agricultural production, and private households were also excluded from the frame.

The establishment-based file was consolidated into a company-based file. In addition, the 4-digit Standard Industrial Classification (SIC) codes for each company were recoded into ACES categories. The 80 ACES categories consisted of either 3-digit SICs or combinations of 3-digit SICs. The ACES sampling frame included approximately two million companies.

## 3. THE L-H METHOD APPLIED TO THE ACES

The universe of companies was classified into two major strata. Stratum I was an arbitrarily defined take-all stratum that consisted of large companies with more than 500 employees and over \$100 million in assets. Stratum I companies were not classified into one ACES industry. For the estimated industry level payroll totals used in the calculation of the industry-level sample sizes, stratum I companies could contribute to more than one ACES industry depending on the number of different ACES industries the companies have payroll in, identified in the SSEL.

Stratum II contained companies that had five or more employees and had less than 500 employees. Stratum II companies were classified into one industry, even if engaged in more than one activity. Each company had frame information available for each of the ACES industries the company had activity in. However, the company's payroll contributed only to estimated total payroll for the industry that the company was classified in. Subsequently, within stratum II, for each ACES industry category, three size strata were created based on total company annual payroll using the L-H method.

A concern with the sample design is the result of companies being misclassified due to the measure of size being used. We classified each stratum II company into its highest payroll industry; however, companies self-report their capital expenditures into ACES industries on the ACES questionnaire. Companies may report in multiple



industries. If too many companies self-report into industries other than where they were classified, then control on the reliability of the estimates is lost.

A similar concern is that the variation in payroll is not the same as the variation in expenditures. Since sample size is directly related to the variance, sample sizes may be different than what is really required. Therefore, since the correlation between payroll and expenditures is not high, the chances that reliability constraints will be met will diminish.

The application of the L-H method to the ACES 1992 preliminary survey sample design involved splitting stratum II into one take-all size stratum and two take-some size strata for each ACES industry. The boundaries were derived for each industry by taking the partial derivative of the sample size with respect to a boundary while fixing the other boundary. However, in practice, we allowed both boundaries to move simultaneously. This results in an iterative process of minimizing the sample size for each industry subject to c.v. constraints. Within stratum II for each ACES industry and assuming Neyman Allocation (Detlefsen and Veum 1991), the sample size equation that is minimized is,

$$n = n_{TA} + \frac{N \left( \sum_{j=1}^2 W_j S_j \right)^2}{\frac{cv^2 Y^2}{N} + \sum_{j=1}^2 W_j S_j^2}, \quad (1)$$

where,  $n_{TA}$  is the number of companies in the take-all size stratum within stratum II defined by the L-H method,  $N$  is the number of stratum II companies in the ACES industry of interest,  $W_j = N_j/N$  is the stratum proportion,  $N_j$  is the number of stratum II companies for size stratum  $j$ ,  $cv$  is the desired coefficient of variation for the ACES industry of interest,  $Y$  is the total payroll for stratum I and II for the ACES industry of interest defined by,

$$Y = \sum_{k=1}^{N_I} y_k + \sum_{j=1}^3 \sum_{i=1}^{N_j} y_{ji},$$

$N_I$  is the number companies in stratum I, and  $S_j$  is the standard deviation of payroll from the SSEL for size stratum  $j$  in stratum II defined by,

$$S_j = \sqrt{\frac{\sum_{i=1}^{N_j} (y_{ji} - \bar{Y}_j)^2}{N_j - 1}},$$

where,  $y_{ji}$  is the payroll value of company  $i$  of size stratum  $j$  for the ACES industry of interest, and  $\bar{Y}_j$  is the mean of payroll for size stratum  $j$ .

The reliability level for each industry was an expected c.v. of 5% on payroll. It was not known, however, what standard errors would result for capital expenditures, as no capital expenditures data exist for the frame records. Companies responding in ACES industries different from the ones they contributed to in the sample design also caused the c.v.'s to fluctuate. The total number of companies selected for the ACES 1992 preliminary survey was 11,194, consisting of 1,500 stratum I companies and 9,694 stratum II companies.

#### 4. CONVERGENCE INTO NEIGHBORHOODS

One of the problems with the L-H method is that it sometimes takes a large number of iterations before the boundaries converge; sometimes they never converge. Generally after just a few iterations, a large proportion of the improvement in the sample size has already occurred. Our goal was to be able to implement stopping rules so that when an area around a local minimum is reached, we can stop processing. This prompted our use of contour plots in analyzing the effect the boundaries have on the resulting sample size. It also allowed us to get a graphical view of the neighborhoods around the local minima. We will use two distributions to illustrate the benefits of reviewing contour plots.

##### 4.1 Non-Skewed Distribution

The first example is a non-skewed distribution from Schneeberger's paper. This distribution is symmetric at  $x = 1$  as shown in Figure 1.

$$f(x) = \begin{cases} 0 & x \leq 0 \\ 2x & 0 < x \leq 0.5 \\ 2(1-x) & 0.5 < x \leq 1 \\ 2(x-1) & 1 < x \leq 1.5 \\ 2(2-x) & 1.5 < x \leq 2 \\ 0 & 2 < x \end{cases}$$

Schneeberger's objective was to find boundaries for three take-some strata using a gradient method. Using the objective function of  $z = (\sum W_h \sigma_h)^2$ , the results attained are listed in Table 1.

**Table 1**  
Optimum Boundaries for Non-Skewed Distribution

	$b_1$	$b_2$	Optimum Point
(2a)	.50241	1.03985	Minimum
(2b)	.70910	1.29090	Saddle Point
(2c)	.96015	1.49759	Minimum

Source: Schneeberger (1979).

**Table 2**  
L-H Boundaries for Three Take-Some Strata for Non-Skewed Distribution

N	Starting Method	1st Iteration			Iteration Within 5% of Sample Size				Final Iteration			
		$b_1$	$b_2$	$n$	$b_1$	$b_2$	$n$	iter.#	$b_1$	$b_2$	$n$	iter.#
50	$N_1 = N_2 = N_3$	.59	1.41	10.89	.66	1.34	9.98	2	.70	1.31	9.77	4
100	$N_1 = N_2 = N_3$	.59	1.41	12.60	.66	1.34	10.91	2	.70	1.30	10.55	5
200	$N_1 = N_2 = N_3$	.59	1.41	13.42	.66	1.34	11.43	2	.71	1.29	10.99	6
1000	$N_1 = N_2 = N_3$	.59	1.41	13.85	.66	1.34	11.75	2	.71	1.29	11.37	7
5000	$N_1 = N_2 = N_3$	.59	1.41	14.12	.66	1.34	11.84	2	.71	1.29	11.45	9
50	Dalenius-Hodges	.70	1.40	10.09	.70	1.40	10.09	1	.77	1.37	9.63	4
100	Dalenius-Hodges	.70	1.40	10.90	.84	1.40	10.14	7	.93	1.47	9.65	13
200	Dalenius-Hodges	.70	1.40	11.42	.83	1.40	10.44	7	.95	1.49	9.96	17
1000	Dalenius-Hodges	.70	1.40	11.86	.86	1.42	10.67	8	.96	1.50	10.27	23
5000	Dalenius-Hodges	.70	1.40	11.95	.86	1.42	10.74	8	.96	1.50	10.34	28
50	Off Line	.50	1.30	10.87	.57	1.20	9.43	3	.55	1.11	9.11	6
100	Off Line	.50	1.30	11.95	.57	1.18	10.04	3	.53	1.07	9.65	8
200	Off Line	.50	1.30	12.64	.56	1.14	10.28	4	.51	1.05	9.96	12
1000	Off Line	.50	1.30	13.24	.56	1.14	10.59	4	.50	1.04	10.27	18
5000	Off Line	.50	1.30	13.37	.56	1.14	10.67	4	.50	1.04	10.34	24

We generated five datasets of different sizes (*e.g.*,  $N = 50, 100, 200, 1000$ , and  $5000$ ) using the formula,  $F(x) = (j - 1/2)/N$ . For this example, we adapted the L-H method to construct three take-some strata and no take-all stratum in order to compare our results with the results in the Schneeberger paper. With our application of estimating totals, when minimizing the sample size subject to a c.v. = 0.05, the L-H method ran for each of the five population sizes using three different starting techniques. The results are given in Table 2.

There are three main points from the information in Table 2. First, the algorithms convergence depends on the population size. The underlying theory of the L-H method is based on continuous distributions. Our examples and any survey application has discrete data from finite populations. It is also apparent that as  $N$  gets larger, the resulting boundaries get closer to where the minimum is under an infinite population size. Figure 2 shows the roughness of the sample size surface when  $N$  is small (*i.e.*,  $N = 50$ ). The resulting surface illustrates the saddle in three dimensions in Figure 2. In this graph, the axes are the lower and upper boundaries and the surface is the resulting sample sizes. This graph shows the saddle-point, the two local minima, and it also gives a picture of the magnitude of the sample size reductions as a result of shifting the boundaries. In contrast, Figure 3 shows the smoothness of the surface when  $N$  is large (*i.e.*,  $N = 5000$ ). From this, it seems that the roughness of the sample size surface and consequently the population size has an effect on where the boundaries converge.

The second point of this example reemphasizes that the ending boundaries are dependent on the starting

boundaries. For this example, Schneeberger describes that with a starting point symmetric to  $x = 1$ , where  $b_1 = 1 - \lambda$  and  $b_2 = 1 + \lambda$  ( $0 < \lambda < 1$ ) which defines the line  $b_2 = 2 - b_1$ , the gradient method moves the gradient along the line  $b_2 = 2 - b_1$  into the saddle-point. When we set the starting boundaries on this line, which occurred when we started with the condition  $N_1 = N_2 = N_3$ , the L-H method also converged to the saddle point (see Table 1). With starting boundaries from the Dalenius-Hodges method, which are not on the line in the case where  $b_2 > 2 - b_1$ , the L-H method converged to a minimum (2c). The Dalenius-Hodges method works well in this example because of the three take-some strata. With starting boundaries which are not on the line in the case where  $b_2 < 2 - b_1$  (specifically,  $b_1 = .5$  and  $b_2 = 1.3$ ), the L-H method converges to a different minimum (2a). This problem is not unique to the L-H method, as Schneeberger points out that the gradient method's resulting boundaries are also dependent of the starting boundaries.

The third point of this example is that there seems to be relatively large reductions in sample size in the first few iterations and then there are several iterations where there are small reductions in sample size. Results are shown in Table 2 from the iteration in which the algorithm produced a sample size within 5% of the final sample size. This implies that the L-H algorithm quickly goes to a neighborhood around an optimal boundary. While close to an optimal sample size, there seems to be a wide range of boundary points resulting in a small range of sample sizes. The point is that stopping rules can save computing time while not relinquishing any real reduction in sample size, since sample size is in integer values.



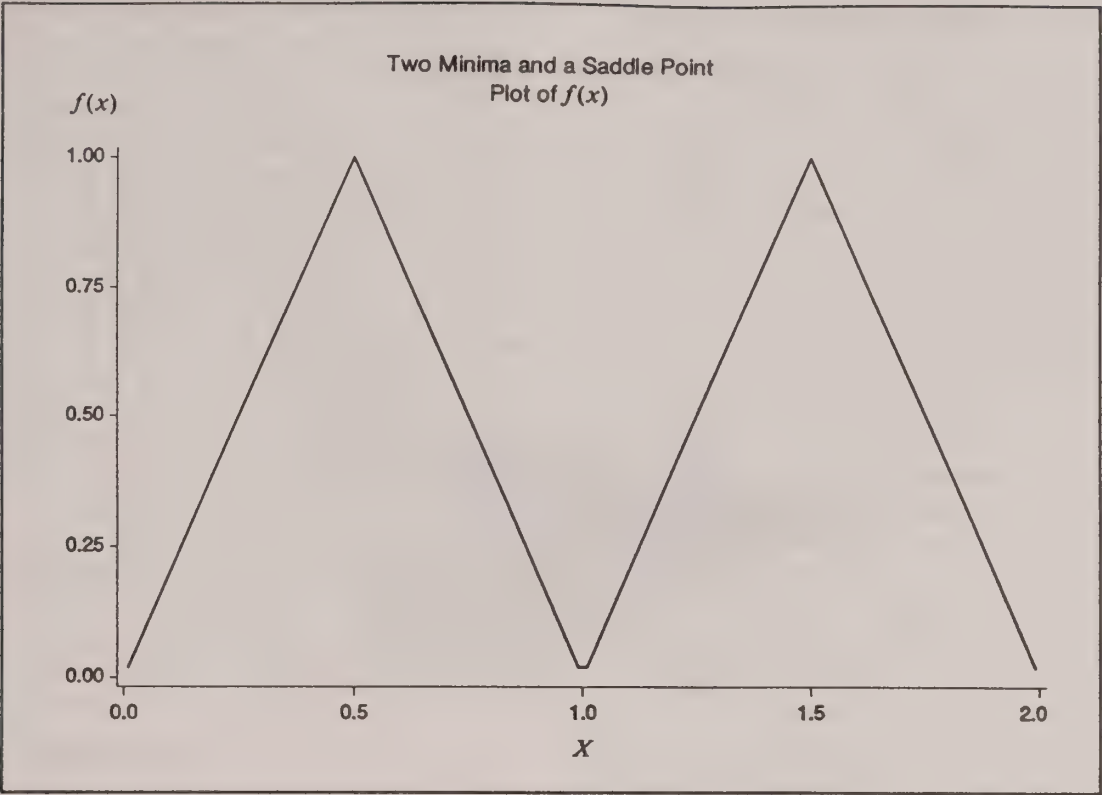


Figure 1. Graph of non-skewed distribution.

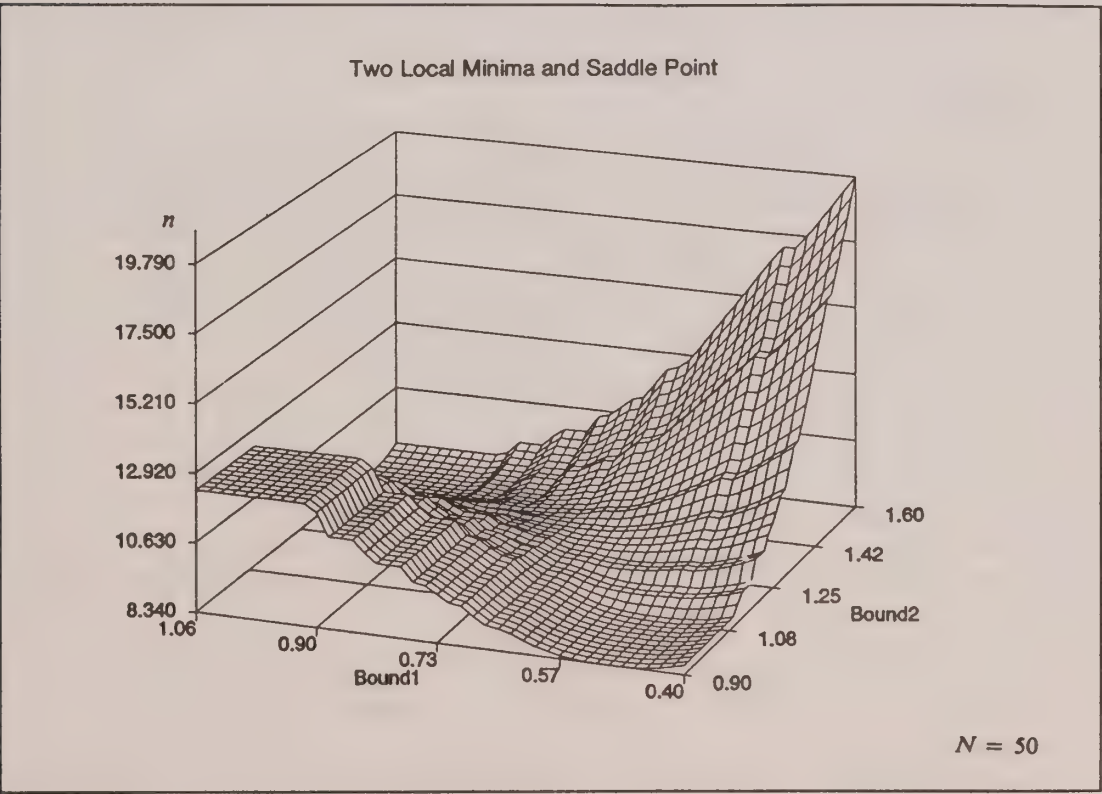


Figure 2. Sample size surface for non-skewed distribution ( $N = 50$ ).

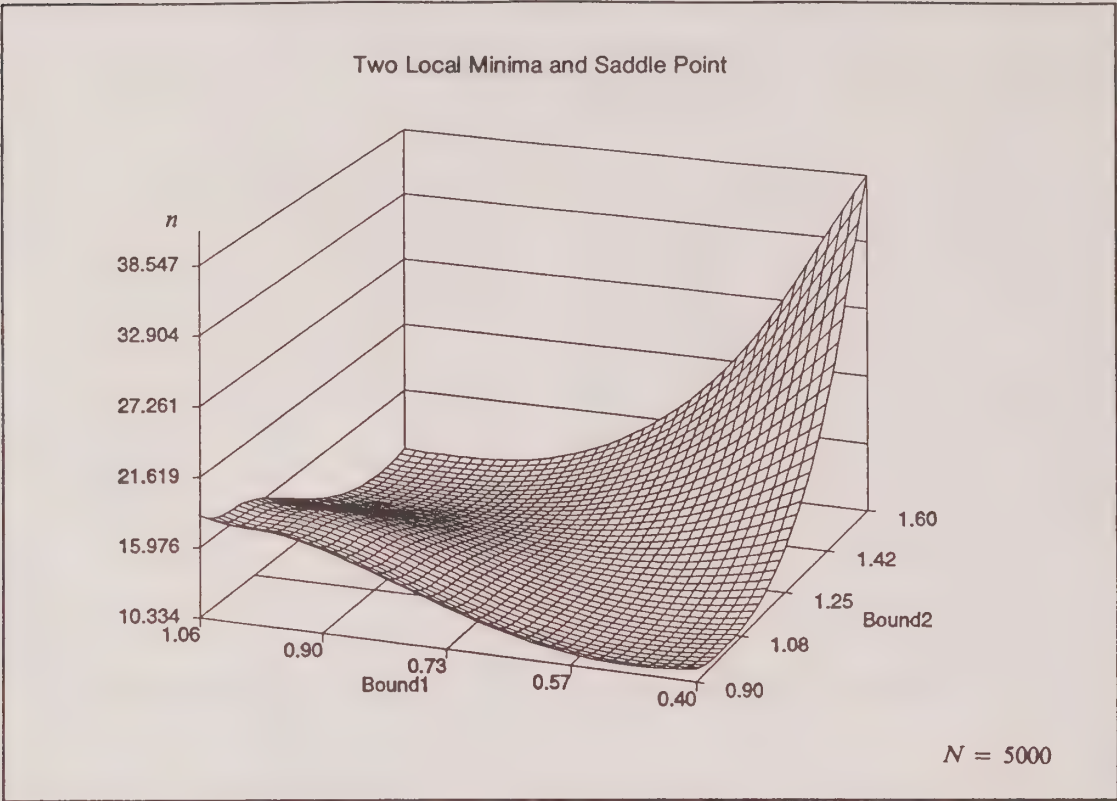


Figure 3. Sample size surface for non-skewed distribution ( $N = 5000$ ).

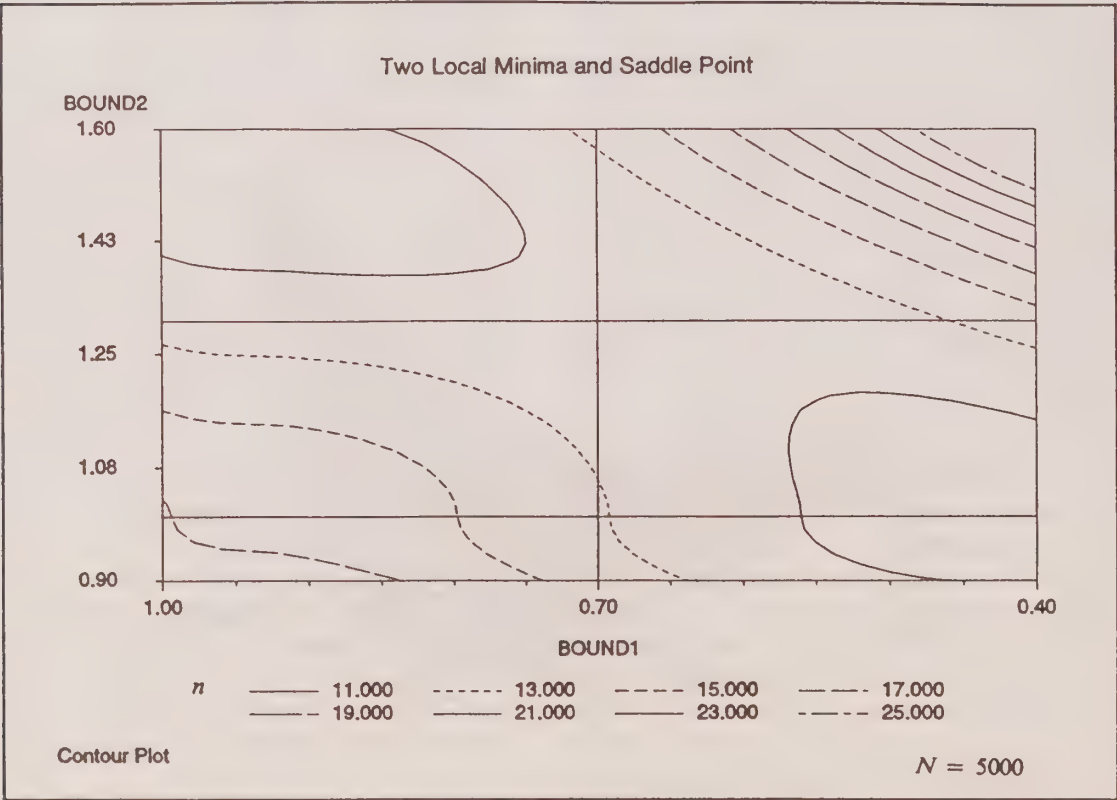


Figure 4. Contour plot for non-skewed distribution ( $N = 5000$ ).



A contour plot of the surface shown in Figure 3 is given in Figure 4. Again, the axes are the lower and upper boundaries and the surface is defined by the resulting sample size. The lines in the plot represent a sample size value. The space between the lines gives an area that contains a range of sample size values. For example, a solid line represents a sample size of 11 and a series of short dash marks represents a sample size of 13. The area in between the solid line and the line of short dash marks contains sample sizes in the range of 11 to 13. This contour plot shows a marginal improvement in the sample size by illustrating that when an area around the bottom of the surface is reached, moving on is unnecessary. At this point, most of the improvement on the sample size from iteration to iteration is less than a value of one. It becomes apparent that after the first few iterations, the improvement of the sample size from iteration to iteration reduces quickly. For instance, in Table 2, where  $N = 5000$  and where the Dalenius-Hodges method was used for the starting boundaries, the first eight iterations accounted for 74% of the total reduction in the sample size from iteration 1 to the 28th and final iteration.

#### 4.2 A Skewed Distribution

Economic data are usually highly skewed and therefore it is more appealing to have a take-all stratum. The next example comes from the Pareto distribution, which is a very typical distribution of economic universes, where there are a large number of small companies and a small number of large companies.

The Pareto distribution function is defined as  $F(x) = 1 - 1/(1 + x)^b$ ,  $0 \leq x < \infty$ . From this we again generated five datasets of different sizes using the formula  $F(x) = (j - 1/2)/N$ . We let the values of  $b$  change as the population size changed. This was done so as to keep the upper tail of the finite discrete distribution roughly the same proportion to the entire population for each population size. To do so, the parameter  $b$  was chosen in such a way that about 90% of the total sum could be accounted for in the top 20% of all possible sampling units. Since the datasets contain a finite number of discrete values there was no problem deriving variances of different strata when values of  $b$  were less than 2.

Table 3 gives the L-H results for different population sizes and starting points. The first group uses starting values which yield equal stratum populations ( $N_1 = N_2 = N_3$ ). The second group uses the Dalenius-Hodges method to obtain all initial boundaries. The third group obtains starting boundaries by first using a method for determining the take-all boundary as presented by Hidioglou (1986) and uses the Dalenius-Hodges method for the other boundary. Again it can be observed that the sample size surface given strata boundaries is much more choppy for smaller population sizes (see Figure 5). For example, when  $N = 50$  and  $b_1$  is fixed, there was only one sample size when  $b_2$  varied between 11.8 and 14.7. This is because there were no values within this range in the population. As the population size increases, the data values are closer together, and the sample surface becomes very smooth (see Figure 6).

Table 3  
L-H Boundaries for Skewed Distribution (one take-all stratum, two take-some strata)

N	Starting Method	1st Iteration					Iteration Within 5% of Sample Size					Final Iteration					
		b	b <sub>1</sub>	b <sub>2</sub>	n <sub>TA</sub>	n	b <sub>1</sub>	b <sub>2</sub>	n <sub>TA</sub>	n	iter.#	b	b <sub>1</sub>	b <sub>2</sub>	n <sub>TA</sub>	n	iter.#
50	N <sub>1</sub> = N <sub>2</sub> = N <sub>3</sub>	.80	.63	2.81	17	17.2	1.66	10.20	7	9.6	5	.80	2.44	11.81	7	9.4	9
100	N <sub>1</sub> = N <sub>2</sub> = N <sub>3</sub>	.90	.56	2.33	34	34.3	1.61	10.29	11	15.8	5	.90	2.58	12.44	10	15.1	12
200	N <sub>1</sub> = N <sub>2</sub> = N <sub>3</sub>	.90	.56	2.36	67	67.2	2.35	17.04	15	21.8	6	.90	3.61	20.46	13	20.9	13
1000	N <sub>1</sub> = N <sub>2</sub> = N <sub>3</sub>	1.00	.50	2.00	333	334.2	3.35	30.58	32	53.0	7	1.00	4.93	36.32	27	51.3	18
5000	N <sub>1</sub> = N <sub>2</sub> = N <sub>3</sub>	1.05	.47	1.85	1665	1667.2	4.67	64.33	62	113.5	7	1.05	7.39	79.38	50	108.8	22
50	Dalenius-Hodges	.80	1.25	8.04	9	10.5	1.76	10.37	7	9.5	3	.80	2.44	11.81	7	9.4	6
100	Dalenius-Hodges	.90	1.39	8.98	13	16.6	1.62	10.16	11	15.8	2	.90	2.58	12.44	10	15.1	9
200	Dalenius-Hodges	.90	1.82	11.66	20	24.3	2.45	17.29	15	21.7	3	.90	3.61	20.46	13	20.9	10
1000	Dalenius-Hodges	1.00	2.37	17.28	55	65.6	3.15	29.70	33	53.5	3	1.00	4.93	36.32	27	51.3	15
5000	Dalenius-Hodges	1.05	3.09	26.27	155	175.0	4.98	66.28	60	112.3	4	1.05	7.39	79.38	50	108.8	19
50	Hidiroglou 1986	.80	.94	6.50	10	11.3	1.58	10.02	7	9.6	3	.80	2.44	11.81	7	9.4	7
100	Hidiroglou 1986	.90	.74	6.17	17	19.6	1.66	10.38	11	15.8	4	.90	2.58	12.44	10	15.1	11
200	Hidiroglou 1986	.90	1.39	9.55	24	27.2	2.50	17.58	14	21.5	4	.90	3.61	20.46	13	20.9	10
1000	Hidiroglou 1986	1.00	2.02	15.13	62	71.3	3.34	30.54	32	53.0	4	1.00	4.93	36.32	27	51.3	15
5000	Hidiroglou 1986	1.05	3.24	28.72	142	164.1	5.11	67.05	59	112.0	4	1.05	7.39	79.38	50	108.8	19

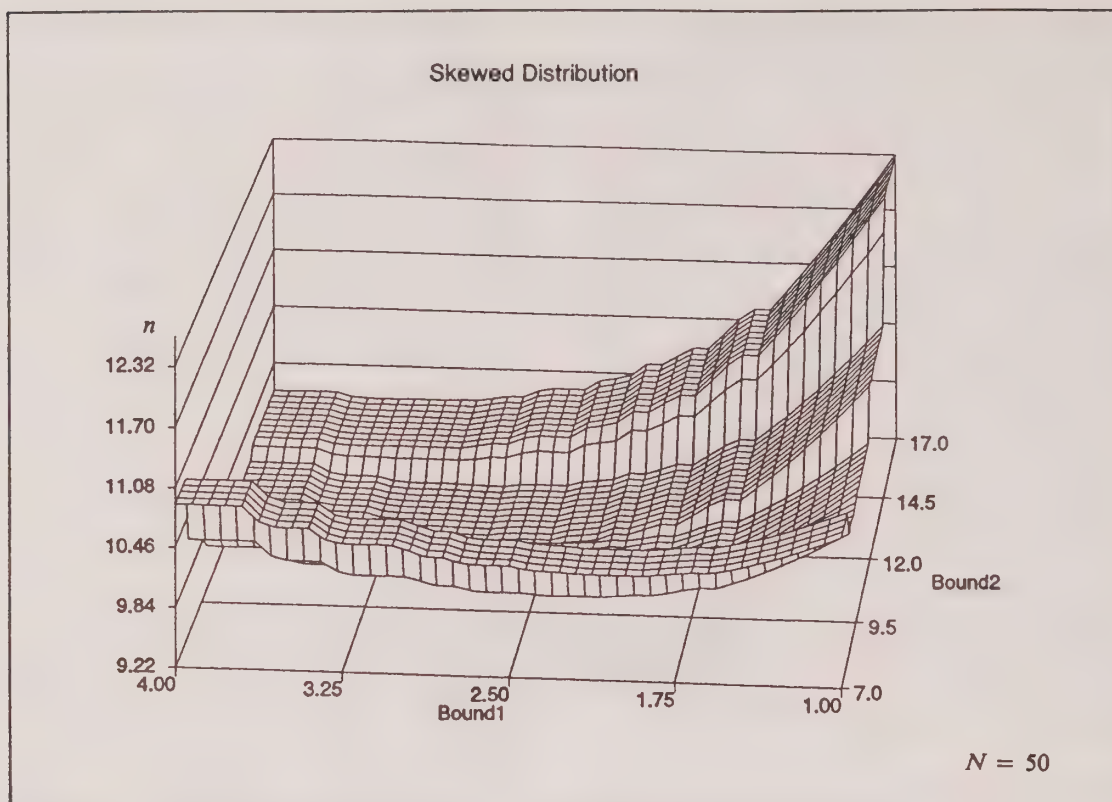


Figure 5. Sample size surface for skewed distribution ( $N = 50$ ).

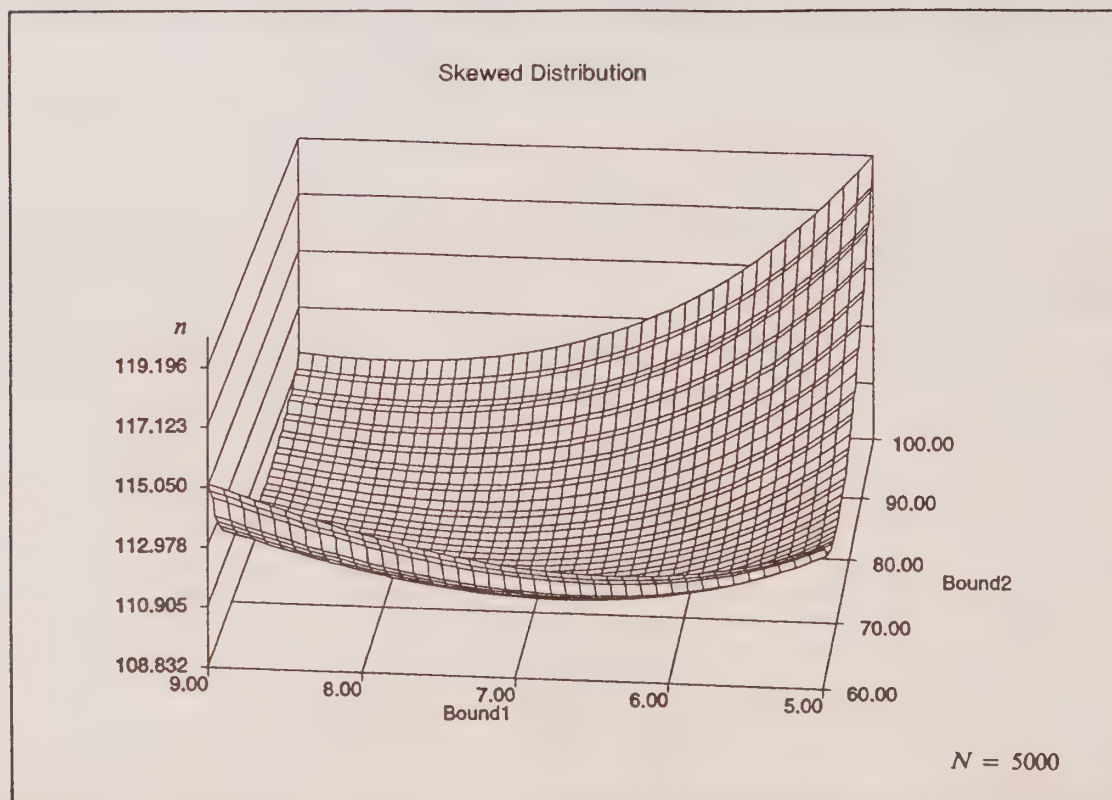


Figure 6. Sample size surface for skewed distribution ( $N = 5000$ ).





Figure 7. Contour plot of skewed distribution ( $N = 50$ ).

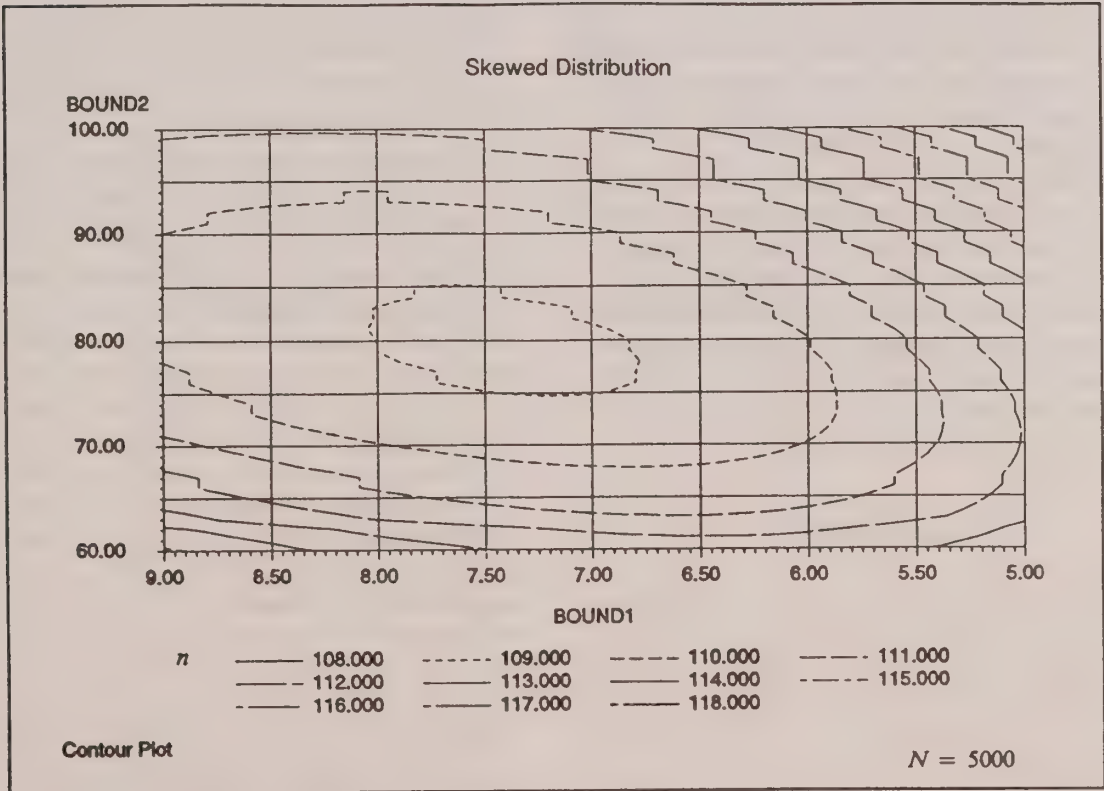


Figure 8. Contour plot of skewed distribution ( $N = 5000$ ).

The contour plot for  $N = 50$  (Figure 7) has erratic shapes defined by straight lines for contour markings. The contour plot for  $N = 5000$  (Figure 8) has almost smooth concentric ellipses for contour markings. It would appear to be a desirable quality for the contour markings to be the same shape and concentric. This would imply that the global minimum is the only local minimum.

The contour plot for  $N = 50$  demonstrated the case where the L-H method didn't converge to optimal boundaries. Since, for this example, we let the L-H program run until it converged the question may arise as to why the L-H method didn't converge to the optimal boundaries. The easiest way to explain this is by viewing Figure 5. We can see that when the population size is small then the sample size surface is not as smooth as in Figure 6. We see several major ridges in Figure 5 that are caused by wide gaps in the skewed discrete data ( $x_{43} = 9.71$ ,  $x_{44} = 11.81$ ,  $x_{45} = 14.79$ ,  $x_{46} = 19.29$ ). This means that for a given  $b_1$ , any value of  $b_2$  between 11.81 and 14.79 would yield the same sample size. When we ran the L-H program for different starting boundaries other than the three listed in Table 3 we came up with the final boundaries as in Table 3 along with other boundaries and their corresponding sample sizes. It appears that the L-H method converges to a low region on one of the major ridges, provided that the region is in the neighborhood of the optimal boundaries. The minimum sample size is 9.22 and the L-H method in Table 3 yielded a sample size of 9.36. The smallest whole integer sample size for each result that meets or exceeds the constraint is 10. Here again we see that the L-H method performs exceptionally well even with discrete distributions that have small population sizes as we see that the boundaries converge within the neighborhood containing the optimal solution.

Another observation to be pointed out is that there is a broad range of values that the boundaries can take on while keeping the integer value of the sample size the same. As the size of the neighborhood expands, the range of boundary values extends as well. It should also be pointed out that even though the range of  $b_1$  values for a given neighborhood is smaller than the range of values for  $b_2$ , there are far more sampling units in the range of  $b_1$  than  $b_2$  because of the skewed distribution.

## 5. SUMMARY

The graphs presented here have shown that a wide range of boundary values result in a small range of sample sizes when in a neighborhood around an optimal value (the bowl shape bottom of the graphs). Any extraordinary improvement on the sample size, *i.e.*, a small marginal gain, might not be worth the extra effort to obtain. This marginal gain may or may not even improve the sample size since the sample size is really an integer and the

marginal gain might only be a small fraction. The L-H method proved very effective in obtaining boundary values in a desired neighborhood around an optimal value, and did it relatively fast.

By measuring the rate of convergence using the sample size instead of boundary values we were better able to determine when a desired neighborhood around an optimal value was reached. This is because boundary values vary greatly in such a neighborhood while sample size (which is of main interest) varies slightly. When the improvement in sample size from iteration to iteration was marginal or nonexistent we immediately terminated the program under the assumption that we reached the desired neighborhood. The following stopping rules are recommended. Stop processing when:

- 1) the difference between the new upper boundary and the previous iteration's upper boundary is less than one. The whole number, one, is used in our case since payroll values are only available to us in whole number values and any shifting of boundaries of a value less than one does not affect any companies;
- 2) the difference between the new lower boundary and the previous iteration's lower boundary is less than one;
- 3) the difference between the new sample size and the previous iteration's sample size is less than a small arbitrary value. We recommend a number less than one since sample sizes are usually rounded up and any fractional improvement on the sample size is negligible. One should be careful when choosing this value since it is possible that the sample size reduction rate may increase from iteration to iteration because the slope of the surface changes;
- 4) the program goes into the 30th iteration. Of course, this is an arbitrary value and may depend on the number of times (industries) one has to apply the L-H method.

Another note is that small population sizes may cause convergence of the boundaries to a point suboptimal, as shown in the examples. Graphs of the sample size surface show a rough surface for small populations and a smooth surface for large populations. It is this rough surface due to the discrete nature of the small population that contributes, in part, to where the L-H method converges.

Another point in conclusion, in our application, the Dalenius-Hodges method assumes that all resulting strata will be sampled. The L-H method is written to construct an analytical take-all substratum. Therefore, the top stratum developed by the Dalenius-Hodges method, when creating the initial boundaries for ACES industries, will be top-heavy since it will not be sampled. Improvements in the sample size were noticed from the Dalenius-Hodges method to the first iteration of the L-H method in this situation. The error that occurs is that the starting boundaries may lead to a local minimum that is not the best solution.



### ACKNOWLEDGEMENTS

The authors are grateful to Michel Hidioglou for useful comments and discussion. We also thank Carol (Veum) Caldwell, Easley Hoy, the referees from *Survey Methodology*, and the Research and Methodology Branch managers of the Manufacturing and Construction Division for helpful comments during review.

### REFERENCES

- COCHRAN, W.G. (1977). *Sampling Techniques*, (3rd Ed.). New York: John Wiley and Sons.
- DETFLESEN, R., and VEUM, C. (1991). Design issues for the retail trade sample surveys of the U.S. Bureau of the Census. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 214-219.
- ECKMAN, G. (1959). An approximation useful in univariate stratification. *The Annals of Mathematical Statistics*, 30, 219-229.
- HESS, I., SETHI, V.K., and BALAKRISHNAN, T.R. (1966). Stratification: A practical investigation. *Journal of the American Statistical Association*, 61, 74-90.
- HIDIROGLOU, M.A. (1986). The construction of a self-representing stratum of large units in survey design. *The American Statistician*, 40, 27-31.
- LAVALLÉE, P., and HIDIROGLOU, M.A. (1988). On the stratification of skewed populations. *Survey Methodology*, 14, 33-43.
- SCHNEEBERGER, H. (1979). Saddle-points of the variance of the sample mean in stratified sampling. *Sankhyā, Series C*, 41, 92-96.
- SETHI, V.K. (1963). A note on optimum stratification of populations for estimating the population means. *American Journal of Statistics*, 5, 20-23.





# A New Method to Reduce Unwanted Ripples and Revisions in Trend-Cycle Estimates From X-11-ARIMA

ESTELA BEE DAGUM<sup>1</sup>

## ABSTRACT

The estimation of the trend-cycle with the X-11-ARIMA method is often done using the 13-term Henderson filter applied to seasonally adjusted data modified by extreme values. This filter however, produces a large number of unwanted ripples in the final or "historical" trend-cycle curve which are interpreted as false turning points. The use of a longer Henderson filter such as the 23-term is not an alternative for this filter is sluggish to detect turning points and consequently is not useful for current economic and business analysis. This paper proposes a new method that enables the use of the 13-term Henderson filter with the advantages of: (i) reducing the number of unwanted ripples; (ii) reducing the size of the revisions to preliminary values and (iii) no increase in the time lag to detect turning points. The results are illustrated with nine leading indicator series of the Canadian Composite Leading Index.

**KEY WORDS:** Trend-cycle; X-11-ARIMA; Turning points; Leading economic indicators.

## 1. INTRODUCTION

The estimation of the trend-cycle with the X-11-ARIMA seasonal adjustment method (Dagum 1980, 1988) as well as the U.S. Bureau of the Census X-11 variant (Shiskin, Young and Musgrave 1967) is done by the application of linear filters due to Henderson (1916). These Henderson filters are applied to seasonally adjusted series where the irregulars have been modified to take into account the presence of extreme values. The length of the filters is automatically selected on the basis of specific values of noise to signal ratios (I/S) being the most commonly chosen the 13-term filter.

The problem of trend-cycle estimation has attracted the attention of several authors, among others, Rhoades (1980); Cholette (1981, 1982); Kenny and Durbin (1982); Castles (1987); Dagum and Laniel (1987); Cleveland, Cleveland, McRae and Terpenning (1990); Wallgren and Wallgren (1990); Gray and Thomson (1990); Findley and Monsell (1990); Scott (1990); and Kenny (1993). Nevertheless, most statistical agencies (excepted the Australian Bureau of Statistics) concentrate their publications on seasonally adjusted series and only very few provide some sort of information on the trend-cycle, usually under the form of graphs.

There are several reasons for limiting the publication of trend-cycle estimates. In the majority of the cases, the seasonally adjusted data are already smooth enough as to be able to provide a clear signal of the short-term trend. But for highly volatile series where further smoothing is required the main objections for trend-cycle estimation are: (1) the size of the revisions of the most recent values (generally much larger than for the corresponding seasonally adjusted estimates) and (2) the presence of short cycles or ripples (9 and 10 months cycles) in the final trend-cycle

curve when the 13-term Henderson filter is applied. On this regard, Kenny (1993) has argued that the presence of ripples in the final estimates of the trend-cycle leads to a large number of false turning points, making the 13-term filter unsuitable for monitoring turning points. He has proposed the use of the 23-term Henderson filter with the object of obtaining a much smoother trend. However, it is well known that this longer filter is sluggish to detect turning points and, hence not useful for current economic and business analysis. For this latter viewpoint, the 13-term filter is preferable but it produces ripples which can be interpreted as false turning points (an unwanted property).

The main purpose of this study is to introduce a new method by which the 13-term Henderson filter can be used with the advantages of: (1) reducing the number of unwanted ripples, (2) reducing the size of the revisions made to the most recent estimates when new observations are added to the series, and (3) not increasing the time lag to detect turning points.

## 2. TREND-CYCLE CASCADE FILTERS

The 13-term Henderson filter is the most often selected and combined with the standard seasonal filters (5- and 7-term moving averages) produces a symmetric cascade filter for final or central values (at least four years from each end of the series) with a gain as exhibited in Figure 1.

Figure 1 also shows the gain functions of other filter convolutions, namely: (1) short seasonal filters with the 9-term Henderson filter and (2) long seasonal filters with the 23-term Henderson filter. It is apparent that cycles of 9 and 10 months (in the 0.08-0.16 frequency band) will not be suppressed by any of the cascade filters, particularly,

<sup>1</sup> Estela Bee Dagum, Faculty of Statistical Sciences, University of Bologna, Via delle Belle Arti 41, (40126) Bologna, Italy.

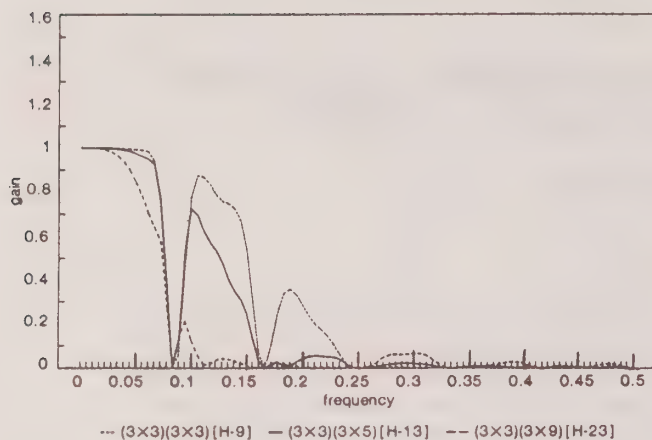


Figure 1. Trend-cycle symmetric cascade filters.

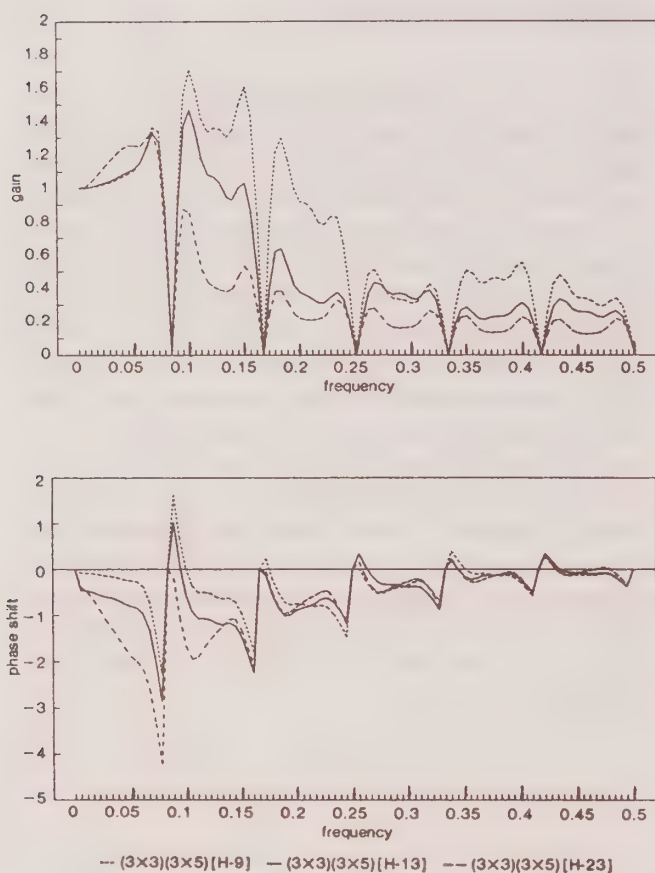


Figure 2. Trend-cycle concurrent cascade filters. Standard seasonal m.a. combined with three Henderson filters.

those using the 9- and 13-term Henderson filters. In fact, the symmetric trend-cycle cascade filter that results from the 9-term Henderson passes about 90% of the power of these short cycles; 72% and 21% are passed by the 13- and 23-term Henderson filters, respectively.

For the concurrent trend-cycle filters which are applied to the last available observation, the peak reached at the frequency band corresponding to 9 and 10 months cycles

is even larger (see Figure 2). Furthermore, all these asymmetric filters introduce phase shift, being near to two months for the 23-term (the largest), one month for the 13-term, and one-half month for the 9-term filter.

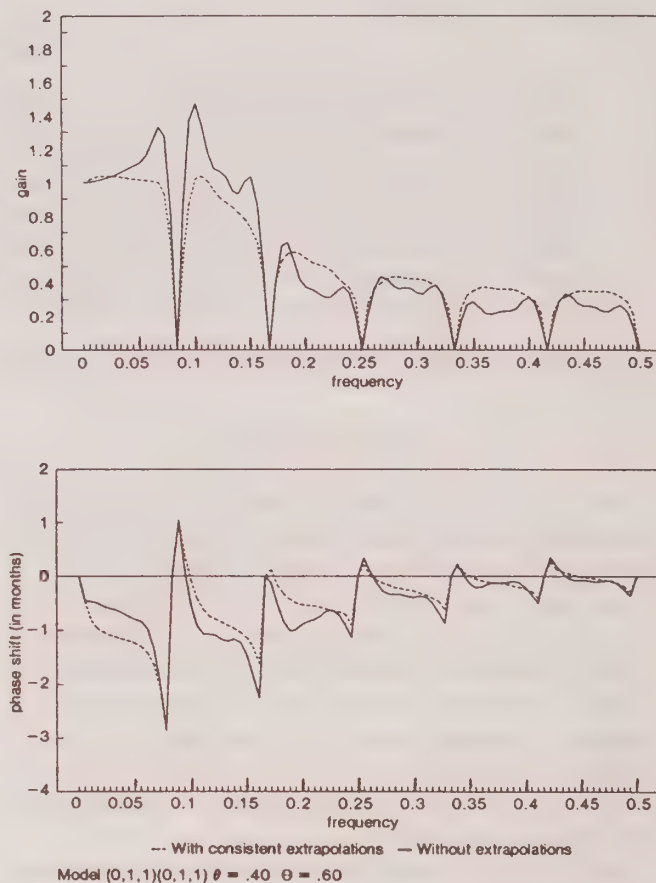


Figure 3. Trend-cycle concurrent cascade filters,  $(3 \times 3)(3 \times 5)[H - 13]$ , with and without ARIMA extrapolations.

Figure 3 shows how the use of ARIMA extrapolations makes the gain of the concurrent cascade filters (using the 13-term Henderson) to resemble the symmetric one although at the expense of a small increase in phase shift. The extrapolations are from an ARIMA model  $(0,1,1)(0,1,1)_s$  where the regular moving average parameter is  $\theta = 0.40$  and the seasonal moving average parameter is  $\Theta = 0.60$ .

Although not shown for space reasons, the gain and phase shift of this trend-cycle concurrent filter fall between the other two combinations.

When ARIMA extrapolations are used, the gain of the concurrent filter converges very fast to that of the final. Dagum and Laniel (1987) show that after three more observations are added to the series, the gain of the asymmetric trend-cycle filter is very close to the symmetric one. The properties of these filters are also extensively discussed in Dagum, Chhab and Chiu (1993, 1996).



The presence of ripples in the final trend-cycle estimates will be produced by the 13-term Henderson filter only if some power is present in the input to the filter at the 0.08-0.16 frequency band. The input to the filter is the seasonally adjusted data with extreme values replaced.

In most empirical cases, the presence of unwanted ripples occurs in periods of high volatility when the observed data are mostly influenced by outliers which can be falsely interpreted as turning points. Although the seasonally adjusted series are modified by extreme values, there is a need for further smoothing which can be done either by applying a longer Henderson filter or by being stricter with the replacement of outliers. Since we want to keep the advantage of a short filter to detect turning points faster, the latter approach is the one followed here.

In the current procedure, the default sigma limits for the replacement of extreme values are  $\pm 1.5$  sigma and  $\pm 2.5$  sigma. Values greater than  $\pm 2.5$  sigma receive a zero weight and those smaller than  $\pm 1.5$  sigma a weight of one (full weight). Values falling within the boundaries are assigned a linearly graduated weight between zero and one.

### 3. A NEW METHOD

The new method here proposed, basically consists of:

(1) extending a smoothed seasonally adjusted series (modified by extreme values with zero weight) with ARIMA extrapolations, and (2) applying the 13-term Henderson filter to the extended series using stricter sigma limits for the identification and replacement of extreme values.

Experimentation with real data showed that the power spectrum of the seasonally adjusted series at the 0.08-0.16 frequency band was drastically reduced only when strict sigma limits such as  $\pm 0.7$  sigma and  $\pm 1.0$  sigma were used. Hence, when applying the 13-term Henderson filter, the trend-cycle curve did not exhibit unwanted ripples while still maintaining its good property of rapid detection of turning points. Under the assumption of normality, these new sigma limits imply that 48% of the irregulars will be modified, 32% will get zero weight and will be replaced by the mean value and 16% will get graduated weights from zero to one.

The extension of the smoothed seasonally adjusted series with ARIMA extrapolations is needed to reduce the size of the revisions for the most recent estimates of the trend-cycle.

The implementation of this new procedure in the context of the X-11-ARIMA and X-11 methods must be done in two steps as follows:

(1) Produce the best seasonally adjusted series selecting appropriate options for the estimation of the components, that is, seasonality, trend-cycle, trading-day variations and Easter effects plus permanent or temporary

priors, if applicable. The seasonally adjusted values are printed in Table D11. The seasonally adjusted series is modified by extreme values with zero weights using the default sigma limits and printed in Table E2. When the estimates of the published seasonally adjusted series for the current year are modified according to some revision practices, then this published revised series should be resubmitted to the X-11-ARIMA program to obtain the corresponding output shown in Table E2.

(2) The output from Table E2 is extended with one year of extrapolations from an ARIMA model. The ARIMA model found adequate with many real series is the (0,1,1) (0,0,1) model. Although the output from Table E2 does not contain seasonality, the seasonal moving average parameter (often of very small value) is needed to correct for some sort of seasonal autocorrelation in the data. The extended series is then run with the X-11-ARIMA program using the Summary Measures option and requesting strict sigma limits ( $\pm 0.7\sigma$  and  $\pm 1.0\sigma$ ) and the 13-term Henderson filter. The new trend-cycle estimates are printed in Table D12.

### 4. EMPIRICAL RESULTS

The new method for trend-cycle estimation is tested with nine leading indicator series of the Canadian Composite Leading Index. In the so called "filtered" version of the Canadian Composite Leading Index published by Statistics Canada, each of the components series as well as the Index itself are smoothed applying to the seasonally adjusted data asymmetric filters based on ARMA models developed by Rhoades (1980). The spectral properties of these ARMA trend-cycle filters are similar to those of the end point of the 9- 13- and 23-term Henderson filters depending on the ARMA model chosen (see Cholette 1982). (Although a comparison with the ARMA filters is not done in this paper, it is likely that the new approach will also give improved results.) Most of the series are highly volatile and all lead at turning points in the business cycle. The series are:

TSE300 Stock Price Index (TSE300)

House Spending Index (HSI)

Money Supply (M1)

Business and Personal Services Employment (BPSE)

Average Workweek in Manufacturing (AWM)

Retail Sales of Furniture and Appliances (RSFA)

Retail Sales of Durable Goods (RSDG)

New Orders for Durable Goods (NODG)

Shipments to Inventories Ratio (SIR).

The advantages of the new procedure versus the currently available in X-11-ARIMA are evaluated as follows.

4.1 Reduction of Ripples in the Final Trend-Cycle Estimates

To calculate the reduction of ripples we first introduce the definition of a turning point within the context of trend-cycle data. A turning point is generally defined as a point in time  $t$  when a series, say  $Y_t$  is larger (smaller) than or equal to the preceding  $k$  and subsequent  $m$  observations of the series. That is,

$$Y_{t-k} \leq \dots \leq Y_{t-1} > Y_t \geq Y_{t+1} \geq \dots \geq Y_{t+m}$$

defines a downturn and

$$Y_{t-k} \geq \dots \geq Y_{t-1} < Y_t \leq Y_{t+1} \leq \dots \leq Y_{t+m}$$

defines an upturn.

From the viewpoint of seasonally adjusted series and trend-cycle data, there is no general consensus for what values of  $k$  and  $m$ , a turning point has occurred. Rhoades (1980) defines a turning point for  $k = 1$  and  $m = 0$ ; Wecker (1979) defines a turning point to be the second of two (or more) successive declines or increases, *i.e.*, for  $k = 2$  and  $m = 2$ ; Zellner, Hong and Min (1991), LeSage (1991) and Pfeffermann and Bleuer (1992) have chosen  $k = 3$  and  $m = 0$ . These definitions do not necessarily correspond to those of cyclical turning points for business cycle analysis but any one can be useful to calculate the number of unwanted ripples as long as two turning points

(a downturn and an upturn) occur within a period of ten months or less. We use here the turning point definition for which  $k = 3$  and  $m = 0$  given the smoothness of the trend-cycle data.

Table 1 shows the number of ripples present in the trend cycle estimates from the standard and the modified 13-term Henderson filter for the period January 1981-December 1993.

Table 1  
Number of Unwanted Ripples in the Trend-Cycle Data  
Using the 13-Term Henderson Filter for the Period  
1981-1993

Series	Standard Procedure	Modified Procedure
NODG	9	2
HSI	8	4
RSDG	8	4
BPSE	8	5
AWM	7	1
SIR	5	1
TS300	4	2
M1	4	2
RSFA	4	0

The results show that the reduction is larger for those series with a large number of ripples and significant in all cases.

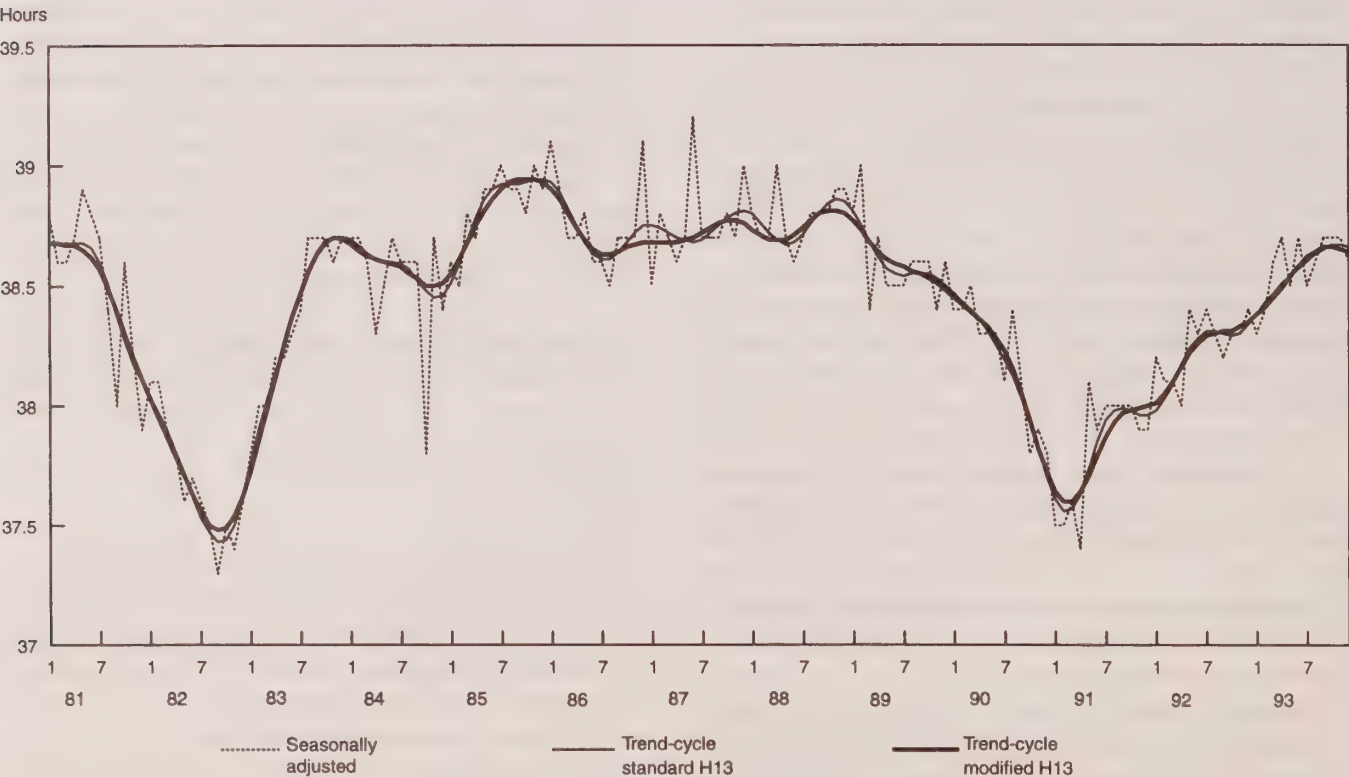


Figure 4. Average work week manufacturing.



Billions of 1981 dollars

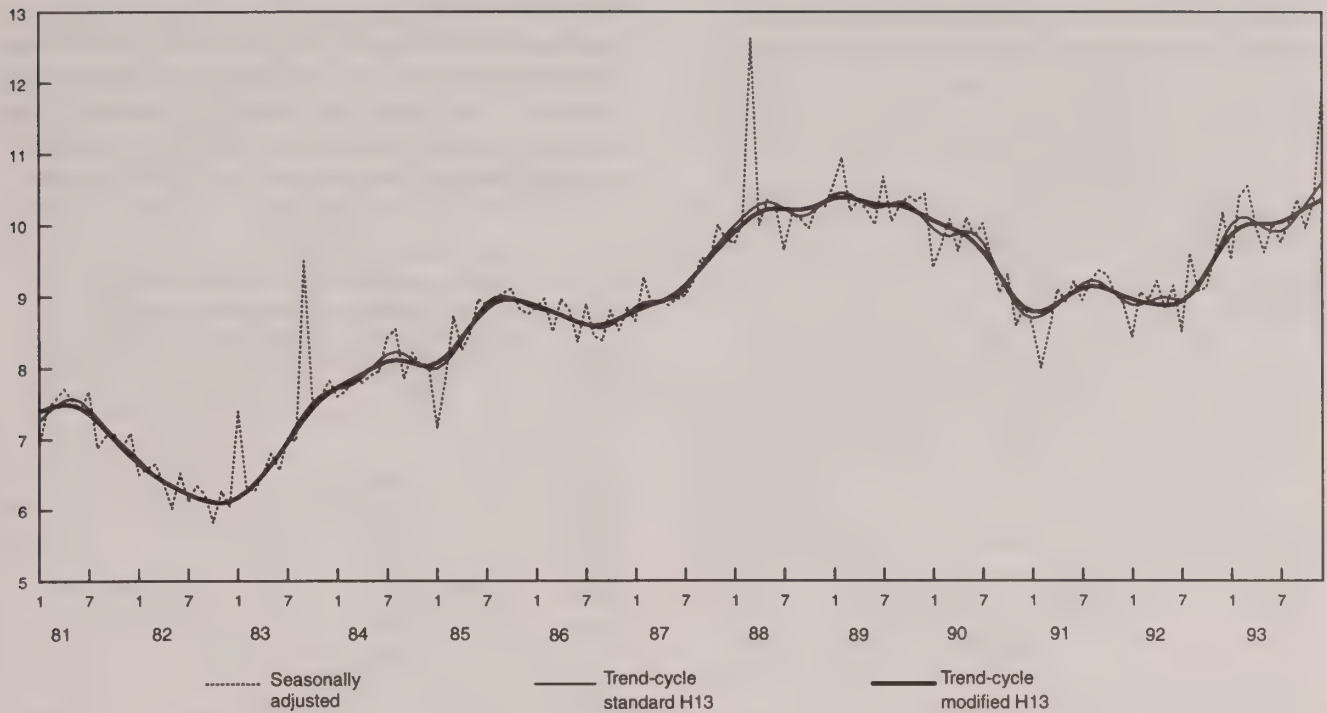


Figure 5. New orders for durable goods.

For illustrative purposes, Figures 4 and 5 for AWM and NODG respectively, exhibit the seasonally adjusted values and the trend-cycle data of both the standard and modified procedures. It is apparent that the new method reduces the ripples in the trend-cycle data with respect to those shown by the standard procedure. In fact, the modified trend-cycle data resembles that of the 23-term Henderson filter but with larger penetration into peaks and troughs of cycles of long duration.

#### 4.2 Turning Point Detection

It is important that the reduction of ripples in the final estimates of the trend-cycle is not achieved at the expense of increasing the lag in detecting turning points which is the main limitation of the 23-term Henderson filter.

To study the revision path of the trend-cycle for any given point in time, the estimates were computed for all end points and previous time points. The revision path of the modified trend-cycle values showed that the identification of cyclical turning points is done with an average lag similar to the standard approach. Depending on the series, the lag was either equal or plus minus one month. For illustrative purposes, Figure 6a. exhibits the revision path of the modified trend-cycle values of New orders for durable goods for the cyclical turning point of February 1991. Successive updates are carried out using data up to March 1991, April 1991 and so on. The turning point is recognized in April, after 2 months whereas it takes

Millions of 1981 dollars

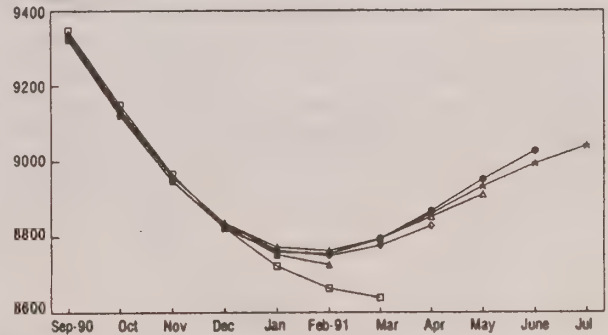


Figure 6a. New orders for durable goods. Trend-cycle modified H13 revisions path.

Millions of 1981 dollars

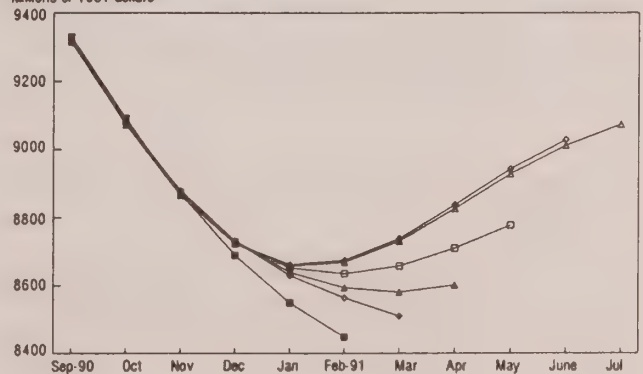


Figure 6b. New orders for durable goods. Trend-cycle standard H13 revisions path.

3 months for the standard procedure as exhibited in Figure 6b. Furthermore, it is shown that successive revisions of the trend-cycle estimates keep generally very close to the final values. The lines which protude, indicating a large revision, can be explained in terms of the underlying data which seem to indicate an increasing decline contradicted by the following values.

Figures 7a. and 7b. for the Average work week in manufacturing reveal that the turning point February-March 1991 is detected three months later by both procedures.

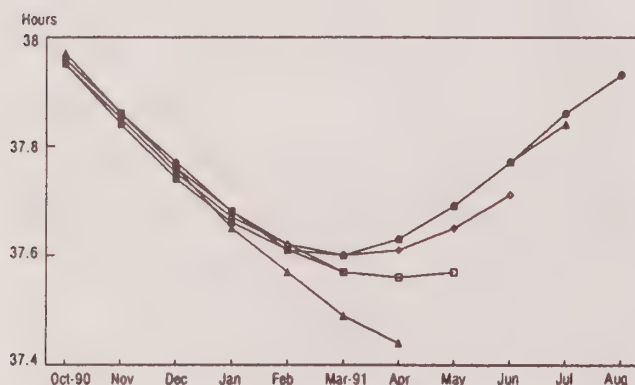


Figure 7a. Average work week manufacturing. Trend-cycle modified H13 revisions path.

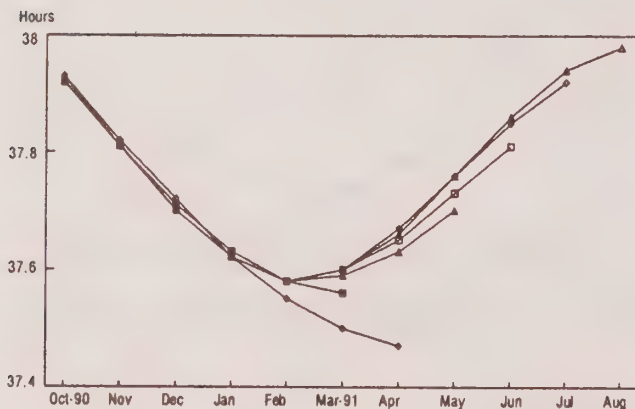


Figure 7b. Average work week manufacturing. Trend-cycle standard H13 revisions path.

#### 4.3 Reduction of Revisions of Concurrent Trend-Cycle Estimates

Another important aspect to take into consideration is to reduce the total revision of the most recent estimate of the trend-cycle which is of preliminary character. Theoretically, the final trend-cycle value is obtained after the series is extended with four years of data but the size of the revisions is negligible after three more months.

Table 2 shows the mean absolute percent revision of the concurrent trend-cycle estimates over a four year period from January 1988 until December 1991. The results indicate that for six of the nine cases analyzed the total revisions of the concurrent trend-cycle values using the modified procedure are much smaller compared to the standard, only for two series they are slightly larger.

**Table 2**  
Mean Absolute Percent Total Revision of  
Concurrent Trend-Cycle  
Values Using the 13-Term Henderson Filter

Series	Standard Procedure (1)	Modified Procedure (2)	Ratio (2)/(1)
NODG	1.55	1.10	0.73
RSFA	0.62	0.47	0.76
RSDG	0.77	0.62	0.80
SIR	0.87	0.70	0.80
AWM	0.13	0.12	0.92
TS300	1.12	1.07	0.95
M1	0.35	0.35	1.00
HSI	2.09	2.20	1.05
BPSE	0.40	0.42	1.05

## 5. CONCLUSION

This paper introduced a new method for trend-cycle estimation which enables the use of the 13-term Henderson filter with the advantages of: (i) reducing the number of unwanted ripples in the final trend-cycle curves, (ii) reducing the size of the revisions to preliminary concurrent values, and (iii) not increase the time lag in turning point detection.

The new method basically consists of extending a smoothed seasonally adjusted series (modified by extreme values with zero weight) with one year of ARIMA extrapolations, and then applying the 13-term Henderson filter using strict sigma limits for the identification and replacement of outliers.

The procedure is illustrated with nine leading indicator series of the Canadian Composite Leading Index and the results are highly satisfactory.

## ACKNOWLEDGEMENTS

This article is based on work supported by Statistics Canada. Views expressed are those of the author and do not necessarily reflect those of Statistics Canada. I am indebted to Marietta Morry and Norma Chhab of the Time Series Research Centre for many stimulating discussions and their collaboration in the work that led to this paper. I am also thankful to an anonymous referee for helpful comments on an earlier version.



## REFERENCES

- CASTLES, I. (1987). A Guide to Smoothing Time Series Estimates of Trend. Catalogue No. 1316.0, Australian Bureau of Statistics.
- CHOLETTE, P.A. (1981). A comparison of various trend-cycle estimators. In *Time Series Analysis*. (O.D. Anderson and M.R. Perryman, Eds). Amsterdam: North-Holland, 77-87.
- CHOLETTE, P.A. (1982). Comparaison de deux estimateurs des cycles économiques. Research Paper No. 82-09-001F, Time Series Research and Analysis Centre, Statistics Canada.
- CLEVELAND, R., CLEVELAND, W.S., McRAE, J.E., and TERPENNING, I. (1990). STL: A seasonal-trend decomposition procedure based on Loess. *Journal of Official Statistics*, 6, 3-33.
- DAGUM, E.B. (1980). The X-11-ARIMA Seasonal Adjustment Method. Catalogue No. 12-564E. Statistics Canada.
- DAGUM, E.B. (1988). The X-11-ARIMA/88 Seasonal Adjustment Method – Foundations and User's Manual. Time Series Research and Analysis Centre, Statistics Canada.
- DAGUM, E.B., and LANIEL, N. (1987). Revisions of trend-cycle estimators of moving average seasonal adjustment methods. *Journal of Business and Economic Statistics*, 5, 177-189.
- DAGUM, E.B., CHHAB, N., and CHIU, K. (1993). Linear properties of the X-11-ARIMA seasonal adjustment method. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*.
- DAGUM, E.B., CHHAB, N., and CHIU, K. (1996). Derivation and properties of the Census X-11 variant and the X-11-ARIMA linear filters. *Journal of Official Statistics*, (forthcoming).
- FINDLEY, D.F., and MONSELL, B.C. (1990). Comment (on Cleveland *et al.* 1990). *Journal of Official Statistics*, 6, 55-59.
- GRAY, A.G., and THOMSON, P.J. (1990). Comment (on Cleveland *et al.* 1990). *Journal of Official Statistics*, 6, 47-54.
- HENDERSON, R. (1916). Note on graduation by adjusted average. *Transactions of the Actuarial Society of America*, 17, 43-48.
- KENNY, P. (1993). Trend presentation. T02919, SMQ, Branch, Central Statistical Office, London, England.
- KENNY, P.B., and DURBIN, J. (1982). Local trend estimation and seasonal adjustment of economic and social time series. *Journal of the Royal Statistical Society, Series A*, 145, 1-41.
- LeSAGE, J.P. (1991). Analysis and development of leading indicators using a Bayesian turning-points approach. *Journal of Business and Economic Statistics*, 9, 305-316.
- PFEFFERMANN, D., and BLEUER, S.R. (1992). Probabilistic detection of nonseasonal turning points in economic time series estimated from sample surveys. Internal report, Methodology Branch, Statistics Canada, Ottawa.
- RHOADES, D. (1980). Converting timeliness into reliability in economic time series or minimum phase shift filtering of economic time series. *Canadian Statistical Review*, 6-13.
- SCOTT, S. (1990). Comment (on Cleveland *et al.* 1990). *Journal of Official Statistics*, 6, 59-62.
- SHISKIN, J., YOUNG, A.H., and MUSGRAVE, J.C. (1967). The X-11 Variant of Census Method II Seasonal Adjustment. Technical Paper No. 15, U.S. Bureau of the Census.
- WALLGREN, B., and WALLGREN, A. (1990). Comment (on Cleveland *et al.* 1990). *Journal of Official Statistics*, 6, 39-46.
- WECKER, W. (1979). Predicting the turning points of a series. *Journal of Business*, 52, 35-50.
- ZELLNER, A., HONG, C., and MIN, C. (1991). Forecasting turning points in international output growth rates using Bayesian exponentially weighted autoregression, time-varying parameter, and pooling techniques. *Journal of Econometrics*, 48, 275-304.





# A Moving Stratification Algorithm

YVES TILLÉ<sup>1</sup>

## ABSTRACT

A general algorithm with equal probabilities is presented. The author provides the second order inclusion probabilities that correspond to the algorithm, which generalizes the selection-rejection method, so that a sample may be drawn using simple random sampling without replacement. Another particular case of the algorithm, called moving stratification algorithm, is discussed. A smooth stratification effect can be obtained by using, as a stratification variable, the serial number of the observation units. The author provides approximations of first and second order inclusion probabilities. These approximations lead to a population mean estimator and to an estimator of the variance of this mean estimator. The algorithm is then compared to a classical stratified plan with proportional allocation.

KEY WORDS: Selection algorithm; Equal probability sampling; Strata.

## 1. INTRODUCTION

When a file is ordered according to an auxiliary variable that is close to the variable of interest, how can a sample be selected using such information? One solution to the problem consists of making a stratified selection. However, making such a selection requires that a delicate problem be resolved, namely subdividing the population into strata. Another simple solution that is both quick and efficient consists of making a systematic selection. The algorithm can be written in a few lines. Moreover, the way in which the file is ordered can be put to good use. However, a systematic selection has one major flaw, namely that estimating the variance of total or mean estimators requires one or several hypotheses concerning the population. It will be shown that there is another simple selection algorithm with which a sample can be drawn in one pass using the file ordering system. For this algorithm, an estimator of the variance of a total or mean estimator is provided, requiring no modelling of the population.

A general selection algorithm providing equal first order inclusion probabilities is presented in section 2. First and second order inclusion probabilities are provided. In section 3, the proposed algorithm is shown to generalize the selection-rejection method so that a simple random sample can be drawn without replacement along with the stratified plan with proportional allocation. Finally, in section 4, the moving stratum method is defined and, in section 5, conclusions are drawn.

## 2. PRESENTATION OF THE GENERAL ALGORITHM

### 2.1 The Algorithm

Let us consider a finite population  $U = \{1, \dots, i, \dots, N\}$ ; we write  $y_1, \dots, y_i, \dots, y_N$ , the  $N$  values assumed

by variable  $y$  for  $N$  observation units of  $U$ . The mean of the values assumed by variable  $y$  for the population is written as

$$\bar{y} = \frac{1}{N} \sum_{i \in U} y_i.$$

A random sample  $s$  of fixed size  $n$  is drawn from this population. The random variables indicating the presence of observation units in  $s$  are written as  $I_i$ ,  $i \in U$ . The first order inclusion probability is written as  $\pi_i = \Pr(i \in s) = E(I_i)$ ,  $i \in U$  and the second order inclusion probability as  $\pi_{ik} = E(I_i I_k)$ ,  $i \neq k \in U$ . The algorithm is very short. It resembles the algorithms of Fan, Fuller and Rezucha (1962), Bebbington (1975), McLeod and Bellhouse (1983) and Sunter (1977, 1986). Only  $N$ ,  $n$  and the  $b_i$ ,  $i = 0, \dots, N - 1$  need to be known. The other variables are working variables.

### General Algorithm

```

j <= 0;
i <= 0;
Repeat for i = 0, ..., N - 1
    u <= a random number with a uniform distribution [0,1];
    if ( (b_i + i)n/N - j ) / b_i > u then
        select record i + 1;
        j <= j + 1;
    otherwise, pass the record i + 1;
    i <= i + 1.

```

At each step,  $j$  represents the number of records already selected and  $i$  the number of records passed (selected or not). For each iteration, a decision is made about selecting the record  $i + 1$ . If the record is selected, it becomes the  $(j + 1)$ -th in the sample. The coefficients  $b_i$ ,  $i = 0, \dots, N - 1$ , are strictly positive real numbers. These

<sup>1</sup> Yves Tillé, Laboratoire de Méthodologie du Traitement des Données, C.P. 124, Université Libre de Bruxelles, avenue Jeanne, 44, 1050 Bruxelles, Belgique, E-mail ytille@ulb.ac.be

quantities must meet certain conditions discussed below if the plan is to be of fixed size or if the units are to be selected with equal probability. The choice of different values for  $b_i$ ,  $i = 0, \dots, N-1$ , will make it possible to generate several special cases of the general algorithm.

If  $b_i$  are strictly positive reals such that  $b_i \leq N-i$ , then the sample size is equal to or smaller than  $n$ . In fact, assuming we have already drawn  $n$  units from the population at step  $i$  and that  $b_i \leq N-i$ , then

$$\frac{(b_i + i)n/N - n}{b_i} = \frac{n}{N} - \frac{n}{b_i} \frac{N-i}{N} \leq \frac{n}{N} - \frac{n}{N-i} \frac{N-i}{N} = 0.$$

It becomes impossible to draw a further unit. It will be assumed in everything that follows that  $b_i \leq N-i$ . Moreover, if  $b_i \leq N-i$ ,  $i = 1, \dots, N-n-1$  and if  $b_i = N-i$ ,  $i = N-n, \dots, N-1$ , the sample is of fixed size  $n$ . Note that these conditions for obtaining a sample of fixed size are sufficient but not necessary.

Three particular cases of the algorithm are examined below. These three cases are defined by three choices of coefficient  $b_i$ ,  $i = 0, \dots, N-1$ . Before examining these particular choices, we will determine the first and second order inclusion probabilities without loss of generality.

## 2.2 First Order Inclusion Probabilities

We write  $n_i$ , the number of units selected after passing  $i$  records. We see immediately that  $n_1, \dots, n_i, \dots, n_N$  is a Markov chain. In fact, we directly derive from the algorithm that

$$\Pr[n_i = j \mid n_1, \dots, n_{i-1}] = \Pr[n_i = j \mid n_{i-1}].$$

The random variables

$$c_i = \frac{(b_i + i)n/N - n_i}{b_i}, \quad i = 0, \dots, N-1,$$

can sometimes assume values greater than 1 or less than 0. Since  $\max(0, n - N + i) \leq n_i \leq \min(i, n)$ , then  $\Pr[0 \leq c_i \leq 1] = 1$  if

$$b_i \geq \begin{cases} \min\left(i \frac{N-n}{n}, N-i\right) & \text{if } n \leq N/2 \\ \min\left(i \frac{n}{N-n}, N-i\right) & \text{if } n > N/2 \end{cases},$$

$$i = 0, \dots, N-1. \quad (1)$$

Again conditions (1) are sufficient but not necessary. We can therefore construct  $b_i$  which do not meet these conditions but which provide  $c_i$  in  $[0, 1]$ . The case dealt with in section 3.2 (stratification) represents one example.

The following example also provides  $c_i$  in  $[0, 1]$  without meeting condition (1): let us consider  $N = 12$ ,  $n = 4$  and  $b_0 = b_1 = b_3 = b_4 = b_6 = 6$ ,  $b_2 = b_5 = 7$ ,  $b_i = N-i$ ,  $i = 12-i$ ,  $i = 7, \dots, 11$ . We have  $c_0 = 1/3$ ,  $c_1 = (7 - 3n_1)/18$ ,  $c_2 = (3 - n_2)/7$ ,  $c_3 = (3 - n_3)/6$ ,  $c_4 = (10 - 3n_4)/18$ ,  $c_5 = (4 - n_5)/7$ ,  $c_6 = (4 - n_6)/6$ ,  $c_7 = (4 - n_7)/5$ ,  $c_8 = (4 - n_8)/4$ ,  $c_9 = (4 - n_9)/3$ ,  $c_{10} = (4 - n_{10})/2$ ,  $c_{11} = (4 - n_{11})$ . We note that  $n_1 \leq 1$ ,  $n_2 \leq 2$ ,  $n_3 \leq 3$ . If  $n_3 = 3$  then  $c_3 = 0$  and therefore  $n_4 \leq 3$ . We then have  $n_5 \leq 4$  and if  $n_5 = 4$  then  $c_5 = 0$  and therefore  $n_6 \leq 4$ . This last comment is true for all  $c_i$  that follow. We therefore note that all  $c_i$  are in  $[0, 1]$  whereas  $b_4 = 6$  does not meet condition (1).

In order to simplify the demonstrations which follow, it will be assumed that

$$\Pr[0 \leq c_i \leq 1] = 1, \quad i = 0, \dots, N-1.$$

We will return to the problem of  $c_i$  values greater than 1 or smaller than 0 later on. If

$$\Pr[0 \leq c_i \leq 1] = 1, \quad i = 0, \dots, N-1,$$

we have

$$E[I_{i+1} \mid n_1, \dots, n_i] = E[I_{i+1} \mid n_i] =$$

$$\frac{(b_i + i)n/N - n_i}{b_i}.$$

It can be shown easily by recursion that if  $\Pr[0 \leq c_i \leq 1] = 1$ ,  $i = 0, \dots, N-1$ ,  $E[n_i] = i n/N$ ,  $i = 0, \dots, N$ . Therefore,

$$\pi_i = E[I_i] = E[n_i] - E[n_{i-1}] = \frac{n}{N}. \quad (2)$$

## 2.3 Second Order Inclusion Probabilities

Four results provided by lemmas 1, 2 and 3 are needed in order to determine second order inclusion probabilities.

**Lemma 1** If  $\Pr[0 \leq c_i \leq 1] = 1$ ,  $i = 0, \dots, N-1$ , then

$$E[n_{i+k} \mid n_i] = (i+k) \frac{n}{N} + \left(n_i - i \frac{n}{N}\right) \prod_{\ell=i}^{i+k-1} \frac{b_\ell - 1}{b_\ell},$$

$$i = 1, \dots, N-1, k = 1, \dots, N-i.$$

This lemma can be demonstrated by recursion if it is assumed to be true for  $k-1$ . Using lemma 1, the following lemma is readily obtained by subtraction:



**Lemma 2** If  $\Pr[0 \leq c_i \leq 1] = 1$ ,  $i = 0, \dots, N-1$ , then

$$E[I_{i+k} | n_i] = \frac{n}{N} - \left(n_i - i \frac{n}{N}\right) \frac{1}{b_{i+k-1}} \prod_{\ell=i}^{i+k-2} \frac{b_\ell - 1}{b_\ell},$$

$$i = 1, \dots, N-1, k = 1, \dots, N-i.$$

It is assumed by convention that an empty product has a value of 1.

**Lemma 3** If  $\Pr[0 \leq c_i \leq 1] = 1$ ,  $i = 0, \dots, N-1$ , then

$$\text{Var}[n_i] = \frac{n}{N} \frac{N-n}{N} \sum_{j=1}^i \prod_{\ell=j}^{i-1} \frac{b_\ell - 2}{b_\ell}, i = 1, \dots, N. \quad (3)$$

The demonstration is provided in the appendix.

Finally, the second order inclusion probability is provided by the following proposition:

**Proposition 1** If  $\Pr[0 \leq c_i \leq 1] = 1$ ,  $i = 0, \dots, N-1$ , then

$$E[I_{i+k} I_{i+1}] = \frac{n^2}{N^2} - \frac{n}{N} \frac{N-n}{N} \frac{1}{b_{i+k-1}}$$

$$\times \left(1 - \frac{1}{b_i} \sum_{j=1}^i \prod_{\ell=j}^{i-1} \frac{b_\ell - 2}{b_\ell}\right) \prod_{\ell=i+1}^{i+k-2} \frac{b_\ell - 1}{b_\ell},$$

$$i = 0, \dots, N-2, k = 2, \dots, N-i. \quad (4)$$

The demonstration is provided in the appendix.

**Corollary 1** If  $\Pr[0 \leq c_i \leq 1] = 1$ ,  $i = 0, \dots, N-1$ , then

$$\pi_{ik} = \frac{n^2}{N^2} - \frac{n}{N} \frac{N-n}{N} \left(1 - \frac{1}{b_{i-1}} \sum_{j=1}^{i-1} \prod_{\ell=j}^{i-2} \frac{b_\ell - 2}{b_\ell}\right)$$

$$\times \frac{1}{b_{k-1}} \prod_{\ell=i}^{k-2} \frac{b_\ell - 1}{b_\ell}, i = 1, \dots, N-1, k > i.$$

## 2.4 The Horvitz-Thompson Estimator and its Variance

The Horvitz-Thompson estimator is the simple sample mean since the first order inclusion probabilities are all equal

$$\hat{y}_\pi = \frac{1}{n} \sum_{i \in S} y_i.$$

If the design is of fixed size, we can use the Yates and Grundy variance formula (1953)

$$\text{Var}[\hat{y}_\pi] = \frac{1}{2N^2} \sum_{i \in U} \sum_{\substack{k \in U \\ k \neq i}} \left(\frac{y_i}{\pi_i} - \frac{y_k}{\pi_k}\right)^2 (\pi_i \pi_k - \pi_{ik}). \quad (5)$$

Since  $\pi_i = n/N$ ,  $i = 1, \dots, N$  and assuming that

$$\gamma_{ik} = 1 - \pi_{ik} \frac{N^2}{n^2},$$

we can write

$$\text{Var}[\hat{y}_\pi] = \frac{1}{N^2} \sum_{i \in U} \sum_{\substack{k \in U \\ k \neq i}} (y_i - y_k)^2 \gamma_{ik}. \quad (6)$$

The variance estimator is provided by

$$\widehat{\text{Var}}[\hat{y}_\pi] = \frac{1}{2N^2} \sum_{i \in S} \sum_{\substack{k \in S \\ k \neq i}} \left(\frac{y_i}{\pi_i} - \frac{y_k}{\pi_k}\right)^2 \frac{\pi_i \pi_k - \pi_{ik}}{\pi_{ik}}. \quad (7)$$

This can be written here as

$$\widehat{\text{Var}}[\hat{y}_\pi] = \frac{1}{2n^2} \sum_{i \in S} \sum_{\substack{k \in S \\ k \neq i}} (y_i - y_k)^2 \frac{\gamma_{ik}}{1 - \gamma_{ik}}.$$

## 3. APPLICATION 1: SIMPLE AND STRATIFIED RANDOM SELECTIONS

### 3.1 Simple Design

The simplest selection algorithm, the selection-rejection method described in Fan, Fuller and Rezucha (1962, method 1), Beddington (1975) and Deville and Grosbras (1987, p. 210), is of course a particular case of the general algorithm. We need only take

$$b_i = N - i, \quad i = 0, \dots, N-1.$$

We always have  $0 \leq c_i \leq 1$ . The first order inclusion probabilities always have a value of  $n/N$ . Calculations for second order inclusion probabilities follow from proposition 1. Assuming  $k > i$ , on the basis of corollary 1, we can find the second order inclusion probabilities of the simple design:

$$\pi_{ik} = \frac{n(n-1)}{N(N-1)}.$$

We also recall some classical results concerning the simple design that we will be using later on. The estimator for  $\bar{y}$  is therefore the mean of the sample

$$\hat{y}_{srs} = \frac{1}{n} \sum_{i \in s} y_i. \quad (8)$$

The variance of this estimator is provided by

$$\text{Var}[\hat{y}_{srs}] = \frac{\sigma_y^2}{n} \frac{N - n}{N - 1} \quad (9)$$

where

$$\sigma_y^2 = \frac{1}{N} \sum_{i \in U} (y_i - \bar{y})^2. \quad (10)$$

An unbiased estimate of this variance is

$$\widehat{\text{Var}}[\hat{y}_{srs}] = \frac{s_y^2}{n} \frac{N - n}{N} \quad (11)$$

where

$$s_y^2 = \frac{1}{n - 1} \sum_{i \in s} (y_i - \hat{y}_{srs})^2. \quad (12)$$

### 3.2 Stratified design

The stratified design can also be defined using the general algorithm. The stratification variable in this case is the serial number of the individual. Let us consider the particular case of a stratified design of  $H$  strata with proportional allocation where all the strata are of the same size. The strata are such that the individuals of a given stratum are adjacent in the data file. It is also assumed that  $N/H$  is an integer. This stratified design is obtained by simply taking

$$b_i = \left\{ (N - i - 1) \bmod \frac{N}{H} \right\} + 1, \quad i = 0, \dots, N - 1.$$

## 4. APPLICATION 2: MOVING STRATIFICATION

### 4.1 The Problem

The file is assumed to be ordered according to an auxiliary variable that is close to the variable of interest. The problem is as follows: how can we draw a random selection that yields a small variance for the Horvitz-Thompson estimator of a mean? Looking at the formulation of the Yates-Grundy variance (5), we see that there are two distinct answers to this question.

The first solution consists of selecting with unequal probabilities using first order inclusion probabilities that are proportional to the variable of interest. If such a selection could be made, all quantities

$$\left( \frac{y_i}{\pi_i} - \frac{y_k}{\pi_k} \right)^2$$

would be zero and therefore the variance would be zero.

The second solution consists of using second order inclusion probabilities. A good selection could be one where  $\pi_{ik}$  are close to  $\pi_i \pi_k$  if  $y_i$  is very different from  $y_k$ . On the other hand, if  $y_i$  is very close to  $y_k$ , we can select second order inclusion probabilities  $\pi_{ik}$  that are clearly smaller than  $\pi_i \pi_k$ . Thus, where quantities

$$\left( \frac{y_i}{\pi_i} - \frac{y_k}{\pi_k} \right)^2$$

would be large (respectively small), quantities  $\pi_i \pi_k - \pi_{ik}$  would be small (respectively large). We would thus have a small variance.

The second solution we have just described is in fact often used. It is the basic idea for stratification. Our objective is to apply this idea to the construction of a sequential selection algorithm that is easy to implement. Such an algorithm could be applied to any file without the need to know anything save the size of the population. It would therefore apply to very large files. We could thus benefit from the information provided by this auxiliary variable like for stratification, without the need to actually subdivide into strata.

### 4.2 The Method

We first define  $M$  the length of the moving stratum within the population.  $M$  represents, in a way, the size of the stratum within the population and is such that  $N/n \leq M \leq N$ . The algorithm of the moving stratum is defined by

$$b_i = \min(M, N - i), \quad i = 0, \dots, N - 1.$$

There is, however, one problem. Quantities  $c_i$  defined by

$$c_i = \begin{cases} \frac{(M + i)n/N - n_i}{M} & \text{if } i \leq N - M \\ \frac{n - n_i}{N - i} & \text{otherwise,} \end{cases}$$

are not always in  $[0, 1]$ .

In fact, let us assume that, before the  $(N - M)$ -th step of the algorithm,  $c_i$  is positive and very close to zero and that through some bad luck the unit  $i$  is nevertheless chosen. In such a case,  $c_{i+1}$  would have a value of  $c_i - (N - n)/(NM)$ .  $c_{i+1}$  can therefore have a negative value but this negative value is always greater than  $-(N - n)/(NM)$ . In fact, if one of the  $c_i$  is already negative, the unit  $i$  is not selected and therefore  $c_{i+1}$  has a value greater than  $c_i$ .

Let us now assume that before the  $(N - M)$ -th step of the algorithm, one  $c_i$  is very slightly smaller than 1 and that nevertheless unit  $i$  is not selected. In such a case,  $c_{i+1}$  would have a value of  $c_i + n/(NM)$ .  $c_{i+1}$  can therefore take on a value greater than 1 but this value greater than 1



is nevertheless always smaller than  $1 + n/(NM)$ . In fact, if one of the  $c_i$  is already greater than 1, the unit  $i$  is always selected and therefore  $c_{i+1}$  has a value smaller than  $c_i$ .

We obtain

$$\Pr\left[-\frac{N-n}{NM} < c_i < 1 + \frac{n}{NM}\right] = 1, i = 0, \dots, N-M. \quad (13)$$

The design is however of fixed size, a result that follows the following proposition:

**Proposition 2** If  $b_i = \min(M, N-i)$ , ( $N/n < M < N$ ),  $0 = 1, \dots, N-1$ , then the design is of fixed size.

The demonstration is provided in the appendix.

Since the  $c_i$  are not always within the interval  $[0,1]$ , we carried out 50 simulations of the moving stratum algorithm for various sample and population sizes. The selected  $N$  population sizes were 100, 500, 2500, 12500, 62500, 312500. The reciprocals of sampling rates ( $N/n$ ) were 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096. We carried out several simulations by varying the size of the moving stratum as follows:  $M = N/n, 2N/n, 3N/n, \dots$ . The simulations seem to indicate that the greater the value for  $M$ , the smaller the probability that a  $c_i$  will fall outside of  $[0,1]$ . As soon as  $M \geq 10N/n$ , for all the simulations that we carried out, the problem was no longer raised. This first result does not imply that the probability that at least one of the  $c_i$  will fall outside of  $[0,1]$  is zero when  $M \geq 10N/n$ . However, it may be said that such a probability would then be very small.

### 4.3 Estimating the Mean and Bias

In examining the results yielded by expression (2) and proposition 1, we get, as a first approximation, a value of about  $\pi_i \approx n/N$  for first order inclusion probabilities. This approximation of inclusion probabilities makes it possible to construct an estimator.

$$\hat{y}_{sm} = \frac{1}{n} \sum_{i \in S} y_i.$$

This estimator is slightly biased since the  $c_i$  are not all exactly within the interval  $[0,1]$ . This bias is

$$B[\hat{y}_{sm}] = \frac{1}{N} \sum_{i \in U} \alpha_i y_i$$

where  $\alpha_i = \pi_i N/n - 1$ . Since the design is of fixed size,  $\sum_{i \in U} \alpha_i = 0$ . We can therefore write the bias in the form of a covariance:  $B[\hat{y}_{sm}] = \sigma_{y\alpha}$  where

$$\sigma_{y\alpha} = \frac{1}{N} \sum_{i \in U} \alpha_i (y_i - \bar{y}). \quad (14)$$

Since the absolute value of a covariance is always equal to or smaller than the product of the two standard deviations, we obtain an upper bound for the absolute value of the bias

$$|B[\hat{y}_{sm}]| \leq \sigma_y \sigma_\alpha$$

where  $\sigma_y$  is defined by (10) and

$$\sigma_\alpha^2 = \frac{1}{N} \sum_{i \in U} \alpha_i^2.$$

The variance of the estimator is of a magnitude that is comparable (for  $N$  and fixed  $n$ ) to the variance of the estimator of the mean in the simple design without replacement. We can therefore write

$$|B[\hat{y}_{sm}]| \leq C_\alpha \sqrt{\text{Var}[\hat{y}_{srs}]}$$

where  $\text{Var}[\hat{y}_{srs}]$  is defined by (9) and

$$C_\alpha = \sigma_\alpha \sqrt{\frac{n(N-1)}{(N-n)}}.$$

We will assume that the bias is negligible when the upper bound of the bias of the estimator  $\hat{y}_{sm}$  is negligible with respect to  $\text{Var}[\hat{y}_{srs}]^{1/2}$ , i.e., when  $C_\alpha$  is small.

Recursively we can calculate the exact value of the  $\Pr[n_i = j]$  since we have

$$\Pr[I_i = 1 | n_i] = \tilde{c}_i, i = 1, \dots, N-M$$

where  $\tilde{c}_i$  has a value of 0 if  $c_i < 0$ ,  $c_i$  if  $0 \leq c_i \leq 1$  and 1 if  $c_i > 1$ . From this result we can derive the exact value of first order inclusion probabilities.

We have calculated (Appendix, Table 1) the values of  $C_\alpha$  for various sample and population (100 – 312500) sizes. The values of  $C_\alpha$  are provided for sizes of moving strata  $M$  equal to  $N/n, 2N/n, 3N/n, 4N/n$  and  $5N/n$ . It can be seen that as soon as the value of the moving stratum is  $2N/n$ ,  $C_\alpha$  never exceeds 0.07. When  $M = 3N/n$ , the coefficient  $C_\alpha$  is expressed in thousandths. According to Cochran (1977, pp. 13-14), the bias is then negligible. The table therefore shows that if  $M \geq 3N/n$ , the bias of the estimator will be negligible at least for the specified sample and population sizes.

However, these results do not imply that the bias of the estimator is large when  $M$  is very small (for example  $M = N/n$ ). The  $C_\alpha$  are bias upper bounds. From expression (14), we see that the bias will be all the greater as the variable of interest correlates with the exact inclusion probabilities. We have shown (Figure 1) the exact inclusion probabilities ( $y$  axis) for  $N$  individuals ( $x$  axis) obtained by using the moving stratification algorithm with the

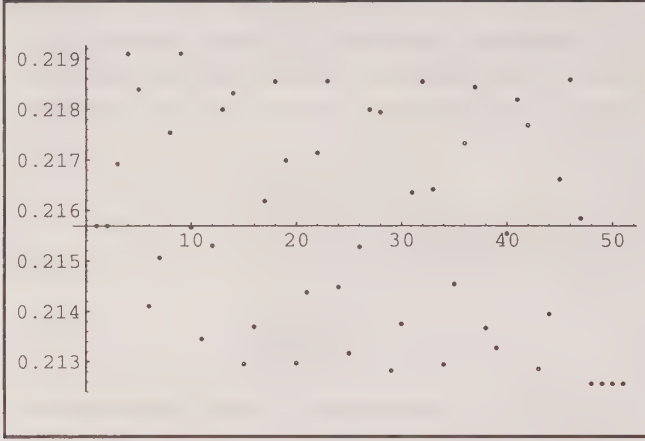


Figure 1. Inclusion probabilities.

parameters  $N = 51$ ,  $n = 11$ ,  $M = N/n$ . This case is obviously very unfavourable. The result is interesting. In this case,  $n/N = 0.215686$ . The inclusion probabilities are distributed on both sides of  $n/N$  with no marked tendency associated with the ordering of the file. In practical terms, the probability can be considered very small that there will be a variable of interest that strongly correlates with the exact inclusion probabilities; as a result, the bias will most often be clearly smaller than the given upper bound.

We could, of course, use the exact inclusion probabilities to establish an estimate. We feel that this is not worthwhile, for two reasons:

- first, because calculating the exact inclusion probabilities requires a significant amount of time,
- second, because the exact first order inclusion probabilities are such that

$$\text{Var} \left[ \sum_{i \in s} \frac{1}{\pi_i} \right] \neq 0.$$

In this case, we have a random Horvitz-Thompson estimator of a constant variable ( $y_k = C$ ). To overcome this problem, an estimate of the mean is usually carried out using Hájek's (1971) ratio. This estimator is also biased.

#### 4.4 Estimating the Variance of the Estimator

Assuming that  $\Pr(0 \leq c_i \leq 1) \approx 1$ , we can also build an approximation of second order inclusion probabilities using corollary 1. Given that  $b_i$  has a value of  $M$  if  $i \leq N - M$  and  $N - i$  otherwise, we obtain the following approximation:

$$\pi_{ik} \approx \frac{n^2}{N^2} (1 - \theta_{ik})$$

where

$$\theta_{ik} = \frac{N - n}{2n} \frac{1}{M - 1} \left\{ 1 + \left( \frac{M - 2}{M} \right)^{\min(i-1, N-M)} \right\} \times \left( \frac{M - 1}{M} \right)^{\max(0, \min(N-M-i+1, k-i))} \quad k > i.$$

Assuming that the first order inclusion probabilities have a value of  $n/N$ , an approximation of the variance of  $\hat{y}_{sm}$  can be obtained:

$$\text{Var}_{app}[\hat{y}_{sm}] = \frac{1}{2N^2} \sum_{i \in U} \sum_{\substack{k \in U \\ k \neq i}} (y_i - y_k)^2 \theta_{ik}. \quad (15)$$

From (15), an estimator of the variance of the estimator of the mean can be obtained:

$$\widehat{\text{Var}}_{app}[\hat{y}_{sm}] = \frac{1}{2N^2} \sum_{i \in s} \sum_{\substack{k \in s \\ k \neq i}} (y_i - y_k)^2 \frac{\theta_{ik}}{1 - \theta_{ik}}. \quad (16)$$

Again, this estimator is biased. In order to assess the magnitude of the bias, we carried out a series of simulations. The results are given in Table 2 in the appendix. We generated populations of size  $N = 400$ . The values assumed by the two variables  $x$  and  $y$  were generated by means of pseudo-random numbers having a bivariate normal distribution with a fixed coefficient of correlation  $\rho$ . The populations were then sorted in terms of the variable  $x$ . The objective was to estimate  $\bar{y}$ .

In these populations, samples of size 64 were selected using the moving stratum method (*sm*), a stratified design with proportional allocation in which the sizes of the strata were all equal (*strat*), as well as a simple design without replacement (*srs*). These three methods are particular cases of the general algorithm and they were implemented using the same random numbers. Simulations were carried out for different values of the moving stratum  $M$  (case: *sm*) and for different numbers of strata  $H$  (case: *strat*). An explanation is provided below for the choices of  $M$  and  $H$ . For each simulation, 200,000 samples were selected.

For each of the simulations, three results are given:

- The means for the simulations of the estimators of the variance of the estimator of the mean, which are expressed as  $E_{sim} \widehat{\text{Var}}(\hat{y})$ . These variance estimators are given by expressions (11) (*srs*) and (16) (*sm*).
- The mean-square errors for the simulations of the estimators of the mean. These quantities are expressed as  $EQM_{sim}(\hat{y}) = E_{sim}(\hat{y} - \bar{y})^2$ .
- The variances of the estimators of the mean. These variances are given by expressions (9) (*srs*) and (15) (*sm*). In the case of the moving stratification, this is of course the proposed approximation.

A careful reading of the results seems to indicate that the variance estimator proposed for the moving stratum algorithm is not affected by a systematic bias no matter what the value for the coefficient of correlation between  $x$  and  $y$ . The results also seem to indicate that the approximate expression given for the variance of the estimator of the mean for the moving stratification is a valid approximation.

#### 4.5 Interest of the Algorithm

Within the class of algorithms defined by the general algorithm, we call the mean horizon of an algorithm the quantity

$$\bar{b} = \frac{1}{N} \sum_{i=0}^{N-1} b_i.$$

For the simple design, we get  $\bar{b}_{srs} = (N + 1)/2$ . For the algorithm of the moving stratum, we have

$$\begin{aligned} \bar{b}_{sm} &= \frac{1}{N} \left\{ \sum_{i=0}^{N-M-1} M + \sum_{i=N-M}^{N-1} (N - i) \right\} \\ &= \frac{M}{N} \left\{ N - \frac{M - 1}{2} \right\}. \end{aligned}$$

Let us now assume that, as described in section 3.2, we select a sample using a design with proportional allocation in which all the strata are of the same size and in which the sizes of  $H$  strata are all equal. In such a design, the mean horizon has a value of

$$\bar{b}_{strat} = \frac{1}{2} \left( \frac{N}{H} + 1 \right).$$

A change in the mean horizon does not fundamentally affect the first order inclusion probabilities. The second order inclusion probabilities, on the other hand, are strongly affected by a change of horizon. In fact, it can easily be seen that the smaller the mean horizon, the smaller the probability of selecting two close individuals. (Two individuals are said to be close if the absolute value of the difference of their serial numbers in the data file is small.) Intuitively, we can expect the moving stratum algorithm to have a stratification effect similar to that of a stratified design with proportional allocation having the same mean horizon, *i.e.*, when

$$\bar{b}_{strat} = \bar{b}_{sm},$$

or in other words, when

$$M = N + \frac{1}{2} - \sqrt{\frac{1}{4} + N^2 \frac{H - 1}{H}}. \quad (17)$$

When  $N$  is large in relation to  $M$ , we have approximately

$$M \approx \frac{2N}{H}.$$

For each series of simulations presented in the Appendix (Table 2), the sizes of the moving strata (case: *sm*) were fixed in terms of the number of strata (case: *strat*) in such a way that the mean horizons of the two designs were identical in terms of expression (17). It is observed that, in such a case, the increased precision (compared to that of the simple design) derived from the moving stratum algorithm is of the same order of magnitude as that derived by means of stratification.

## 5. COMMENTS

The simulations that were carried out clearly show that the moving stratification algorithm yields a stratification effect of the same type as classical stratification with proportional allocation. This algorithm makes it possible to study the delicate problem of subdividing a continuous variable into strata. The estimators of the mean that are proposed are slightly biased. However, as long as  $M \geq 10N/n$ , simulations show that it is extremely rare for at least one of the  $c_i$  to fall outside of  $[0, 1]$ . Moreover, we have shown that even when that probability is not zero, the bias of the estimator that we propose is negligible as long as  $M \geq 3N/n$ .

## ACKNOWLEDGEMENTS

The author wishes to thank Pierre Lavallée for comments made on previous drafts of this paper. The author also wishes to thank an associate editor and a referee who provided numerous constructive comments that significantly improved this paper.



## APPENDIX 1

### Demonstration of the Lemmas and Propositions

#### Demonstration of Lemma 3

$$\text{Var}[n_{i+1}]$$

$$= \text{Var}[n_i] + \text{Var}[I_{i+1}]$$

$$+ 2E\left(E\left\{\left(n_i - i \frac{n}{N}\right)E\left[I_{i+1} - \frac{n}{N} \mid n_i\right]\right\}\right).$$

Since

$$\begin{aligned} & 2E\left[E\left\{\left(n_i - i \frac{n}{N}\right)E\left[\left(I_{i+1} - \frac{n}{N}\right) \mid n_i\right]\right\}\right] \\ &= 2E\left[\left(n_i - i \frac{n}{N}\right)\left(\frac{(b_i + i)n/N - n_i}{b_i} - \frac{n}{N}\right)\right] \\ &= \frac{-2}{b_i} \text{Var}[n_i], \end{aligned}$$

we obtain

$$\begin{aligned} \text{Var}[n_{i+1}] &= \text{Var}[n_i] \frac{b_i - 2}{b_i} + \frac{n}{N} \frac{N - n}{N}, \\ i &= 1, \dots, N - 1. \end{aligned} \quad (18)$$

We then show that (3) verifies the recursion equation (18) and the initial condition given by

$$\text{Var}(n_1) = \frac{n}{N} \frac{N - n}{N}.$$

#### Demonstration of Proposition 1

Case 1:  $i = 0$ . From lemma 2 we immediately get:

$$\begin{aligned} E[I_k I_1] &= E[E[I_k \mid n_1] n_1] \\ &= \frac{n^2}{N^2} - \frac{n}{N} \frac{N - n}{N} \frac{1}{b_{k-1}} \prod_{\ell=1}^{k-2} \frac{b_\ell - 1}{b_\ell}. \end{aligned}$$

Case 2:  $i > 0$ . Using lemma 2, we obtain:

$$\begin{aligned} E[I_{i+k} I_{i+1} \mid n_i = t] \\ &= E[I_{i+k} \mid n_{i+1} = t + 1] E[I_{i+1} \mid n_i = t] \\ &= \left\{ \frac{n}{N} - \left( (t+1) - (i+1) \frac{n}{N} \right) \frac{1}{b_{i+k-1}} \prod_{\ell=i+1}^{i+k-2} \frac{b_\ell - 1}{b_\ell} \right\} \\ &\quad \times \left\{ \frac{n}{N} - \left( t - i \frac{n}{N} \right) \frac{1}{b_i} \right\}. \end{aligned}$$

Which means that

$$\begin{aligned} E[E[I_{i+k} I_{i+1} \mid n_i]] \\ &= E\left\{ \frac{n}{N} - \left( (n_i + 1) - (i+1) \frac{n}{N} \right) \frac{1}{b_{i+k-1}} \prod_{\ell=i+1}^{i+k-2} \frac{b_\ell - 1}{b_\ell} \right\} \\ &\quad \times \left\{ \frac{n}{N} - \left( n_i - i \frac{n}{N} \right) \frac{1}{b_i} \right\} \\ &= \frac{n^2}{N^2} - \frac{1}{b_{i+k-1}} \left\{ \frac{n}{N} \frac{N - n}{N} - \frac{\text{Var}[n_i]}{b_i} \right\} \prod_{\ell=i+1}^{i+k-2} \frac{b_\ell - 1}{b_\ell}. \end{aligned}$$

Lemma 3 thus gives us  $\text{Var}[n_i]$ . We immediately obtain (4).

#### Demonstration of Proposition 2

Using (13), we have

$$\Pr\left[n - M - \frac{n}{N} < n_{N-M} < \frac{N - n}{N} + n\right] = 1.$$

Therefore,

$$\Pr[0 \leq n - n_{N-M} \leq M] = 1.$$

Beginning with step  $N - M$ , the algorithm is a selection-rejection algorithm of the type described in section 3.1. This algorithm yields a sample of exactly  $n - n_{N-M}$  observation units during the final  $M$  steps. Since  $n - n_{N-M} \leq M$ , this operation raises no difficulty and the algorithm is therefore of fixed size  $n$ .

## APPENDIX 2

## Tables, Bias Upper Bounds and Simulations

Table 1

Value of the Bias Upper Bounds  $C_\alpha$ 

N	n	Value of the Coefficient $C_\alpha$				
		$M = \frac{N}{n}$	$M = \frac{2N}{n}$	$M = \frac{3N}{n}$	$M = \frac{4N}{n}$	$M = \frac{5N}{n}$
100	50	0.000000	0.000000	0.000000	0.000000	0.000000
	25	0.057326	0.002610	0.000185	0.000015	0.000001
	12	0.041716	0.002604	0.000235	0.000023	0.000002
	6	0.032227	0.002029	0.000134	0.000005	0.000000
	3	0.023515	0.000645	0.000000		
500	250	0.000000	0.000000	0.000000	0.000000	0.000000
	125	0.129091	0.006002	0.000437	0.000038	0.000004
	62	0.090863	0.005664	0.000534	0.000059	0.000007
	31	0.066891	0.004666	0.000484	0.000059	0.000008
	15	0.048544	0.003586	0.000384	0.000046	0.000006
	7	0.035508	0.002552	0.000215	0.000015	0.000001
	3	0.024046	0.000699	0.000000		
2,500	1,250	0.000000	0.000000	0.000000	0.000000	0.000000
	625	0.289060	0.013495	0.000987	0.000086	0.000008
	312	0.202458	0.012607	0.001190	0.000133	0.000016
	156	0.147113	0.010234	0.001064	0.000130	0.000017
	78	0.105662	0.007742	0.000841	0.000107	0.000015
	39	0.075975	0.005719	0.000634	0.000082	0.000012
	19	0.054525	0.004174	0.000466	0.000060	0.000008
	9	0.039560	0.003014	0.000301	0.000029	0.000002
	4	0.028388	0.001451	0.000034	0.000000	
12,500	3,125	0.646539	0.030208	0.002211	0.000193	0.000018
	1,562	0.452450	0.028177	0.002661	0.000297	0.000036
	781	0.327879	0.022798	0.002371	0.000290	0.000039
	390	0.234114	0.017131	0.001863	0.000238	0.000033
	195	0.166626	0.012500	0.001388	0.000181	0.000026
	97	0.118357	0.008995	0.001009	0.000133	0.000019
	48	0.084217	0.006452	0.000727	0.000096	0.000014
	24	0.060797	0.004689	0.000529	0.000069	0.000010
	12	0.044677	0.003461	0.000377	0.000044	0.000005
	6	0.033727	0.002356	0.000173	0.000008	0.000000
	3	0.024172	0.000712	0.000000		
62,500	3,906	0.732684	0.050942	0.005299	0.000649	0.000087
	1,953	0.522918	0.038250	0.004159	0.000531	0.000074
	976	0.371301	0.027833	0.003092	0.000403	0.000057
	488	0.263300	0.019979	0.002243	0.000295	0.000042
	244	0.186736	0.014259	0.001609	0.000213	0.000031
	122	0.132653	0.010168	0.001150	0.000152	0.000022
	61	0.094601	0.007273	0.000823	0.000109	0.000016
	30	0.067467	0.005207	0.000590	0.000078	0.000011
	15	0.049227	0.003820	0.000427	0.000054	0.000007
	7	0.035847	0.002637	0.000227	0.000016	0.000001
	3	0.024176	0.000713	0.000000		
312,500	4,882	0.829762	0.062191	0.006909	0.000901	0.000128
	2,441	0.587909	0.044596	0.005006	0.000659	0.000095
	1,220	0.416165	0.031758	0.003583	0.000474	0.000068
	610	0.294647	0.022555	0.002551	0.000339	0.000049
	305	0.208743	0.016008	0.001813	0.000241	0.000035
	152	0.147877	0.011356	0.001287	0.000171	0.000025
	76	0.105272	0.008098	0.000918	0.000122	0.000018
	38	0.075422	0.005817	0.000659	0.000087	0.000013
	19	0.054695	0.004238	0.000479	0.000062	0.000009
	9	0.039644	0.003038	0.000305	0.000030	0.000002
	4	0.028427	0.001457	0.000034	0.000000	

Table 2

Results of the Simulations, Simple Design, Stratification and Moving Stratification

$\rho^2$	Plan	Parameters	$E_{sim}\widehat{\text{Var}}\hat{y}$	$\text{Var}\hat{y}$	$EQM_{sim}\hat{y}$
0.0	sm	$M = 18.83N/n$	0.01318	0.01317	0.01301
	srs		0.01317	0.01316	0.01296
	strat	$H = 2$	0.01319	0.01319	0.01318
0.2	sm	$M = 18.83N/n$	0.01210	0.01210	0.01187
	srs		0.01316	0.01316	0.01287
	strat	$H = 2$	0.01172	0.01188	0.01164
0.4	sm	$M = 18.83N/n$	0.01073	0.01073	0.01080
	srs		0.01316	0.01316	0.01320
	strat	$H = 2$	0.00943	0.00929	0.00946
0.6	sm	$M = 18.83N/n$	0.00957	0.00957	0.00954
	srs		0.01315	0.01316	0.01301
	strat	$H = 2$	0.00783	0.00778	0.00774
0.8	sm	$M = 18.83N/n$	0.00839	0.00839	0.00839
	srs		0.01315	0.01316	0.01322
	strat	$H = 2$	0.00630	0.00624	0.00622
1.0	sm	$M = 18.83N/n$	0.00757	0.00757	0.00760
	srs		0.01314	0.01316	0.01319
	strat	$H = 2$	0.00514	0.00508	0.00513
0.0	sm	$M = 8.65N/n$	0.01319	0.01319	0.01317
	srs		0.01317	0.01316	0.01296
	strat	$H = 4$	0.01320	0.01318	0.01316
0.2	sm	$M = 8.65N/n$	0.01107	0.01107	0.01084
	srs		0.01316	0.01316	0.01287
	strat	$H = 4$	0.01080	0.01076	0.01054
0.4	sm	$M = 8.65N/n$	0.00876	0.00876	0.00882
	srs		0.01316	0.01316	0.01320
	strat	$H = 4$	0.00811	0.00793	0.00796
0.6	sm	$M = 8.65N/n$	0.00695	0.00694	0.00688
	srs		0.01315	0.01316	0.01301
	strat	$H = 4$	0.00637	0.00639	0.00632
0.8	sm	$M = 8.65N/n$	0.00484	0.00484	0.00485
	srs		0.01315	0.01316	0.01322
	strat	$H = 4$	0.00402	0.00391	0.00390
1.0	sm	$M = 8.65N/n$	0.00312	0.00312	0.00313
	srs		0.01314	0.01316	0.01319
	strat	$H = 4$	0.00206	0.00197	0.00197
0.0	sm	$M = 4.21N/n$	0.01317	0.01317	0.01316
	srs		0.01317	0.01316	0.01296
	strat	$H = 8$	0.01321	0.01324	0.01325
0.2	sm	$M = 4.21N/n$	0.01067	0.01067	0.01046
	srs		0.01316	0.01316	0.01287
	strat	$H = 8$	0.01055	0.01047	0.01025
0.4	sm	$M = 4.21N/n$	0.00810	0.00809	0.00808
	srs		0.01316	0.01316	0.01320
	strat	$H = 8$	0.00794	0.00789	0.00789
0.6	sm	$M = 4.21N/n$	0.00592	0.00592	0.00588
	srs		0.01315	0.01316	0.01301
	strat	$H = 8$	0.00575	0.00564	0.00561
0.8	sm	$M = 4.21N/n$	0.00344	0.00344	0.00345
	srs		0.01315	0.01316	0.01322
	strat	$H = 8$	0.00315	0.00311	0.00308
1.0	sm	$M = 4.21N/n$	0.00124	0.00124	0.00125
	srs		0.01314	0.01316	0.01319
	strat	$H = 8$	0.00085	0.00079	0.00080

Table 2

Results of the Simulations, Simple Design, Stratification and Moving Stratification – end

$\rho^2$	Plan	Parameters	$E_{sim}\widehat{\text{Var}}\hat{y}$	$\text{Var}\hat{y}$	$EQM_{sim}\hat{y}$
0.0	sm	$M = 2.11N/n$	0.01319	0.01319	0.01328
	srs		0.01315	0.01316	0.01332
	strat	$H = 16$	0.01315	0.01308	0.01331
0.2	sm	$M = 2.11N/n$	0.01038	0.01036	0.01021
	srs		0.01317	0.01316	0.01334
	strat	$H = 16$	0.01034	0.01034	0.01025
0.4	sm	$M = 2.11N/n$	0.00796	0.00796	0.00792
	srs		0.01316	0.01316	0.01323
	strat	$H = 16$	0.00790	0.00801	0.00794
0.6	sm	$M = 2.11N/n$	0.00572	0.00573	0.00561
	srs		0.01315	0.01316	0.01299
	strat	$H = 16$	0.00568	0.00572	0.00563
0.8	sm	$M = 2.11N/n$	0.00295	0.00294	0.00290
	srs		0.01317	0.01316	0.01325
	strat	$H = 16$	0.00287	0.00288	0.00285
1.0	sm	$M = 2.11N/n$	0.00048	0.00048	0.00048
	srs		0.01317	0.01316	0.01335
	strat	$H = 16$	0.00037	0.00034	0.00034
0.0	sm	$M = 1.09N/n$	0.01325	0.01316	0.01310
	srs		0.01313	0.01316	0.01317
	strat	$H = 32$	0.01201	0.01239	0.01302
0.2	sm	$M = 1.09N/n$	0.01070	0.01062	0.01064
	srs		0.01313	0.01316	0.01316
	strat	$H = 32$	0.00972	0.01018	0.01083
0.4	sm	$M = 1.09N/n$	0.00807	0.00803	0.00811
	srs		0.01315	0.01316	0.01309
	strat	$H = 32$	0.00732	0.00751	0.00803
0.6	sm	$M = 1.09N/n$	0.00538	0.00534	0.00536
	srs		0.01315	0.01316	0.01310
	strat	$H = 32$	0.00484	0.00484	0.00543
0.8	sm	$M = 1.09N/n$	0.00283	0.00281	0.00276
	srs		0.01317	0.01316	0.01283
	strat	$H = 32$	0.00255	0.00276	0.00280
1.0	sm	$M = 1.09N/n$	0.00016	0.00016	0.00017
	srs		0.01317	0.01316	0.01304
	strat	$H = 32$	0.00012	0.00007	0.00011

## REFERENCES

- BEBBINGTON, A.C. (1975). A simple method of drawing a sample without replacement. *Applied Statistics*, 24, 136.
- COCHRAN, W.G. (1977). *Sampling Techniques*. New York: Wiley.
- DEVILLE, J.-C., and GROSBAS, J.-M. (1987). Algorithmes de tirage. In *Les sondages*. Dreesbeke, J.-J., Fichet, B., and Tassi, P. (Eds.). Paris: Economica, 209-233.
- FAN, C.T., MULLER, M.E., and REZUCHA, I. (1962). Development of sampling plans by using sequential (item by item) selection techniques and digital computers. *Journal of the American Statistical Association*, 57, 387-402.
- HÁJEK, J. (1971). Comment on an essay of D. Basu. In *Foundations of Statistical Inference*. Godambe V.P., and Sprott, D.A. (Eds.). Toronto: Holt, Rinehart and Winston.
- MCLEOD, A.I., and BELLHOUSE, D.R. (1983). A convenient algorithm for drawing a simple random sampling. *Applied Statistics*, 32, 182-184.
- SUNTER, A.B. (1977). List sequential sampling with equal or unequal probabilities without replacement. *Applied Statistics*, 26, 261-268.
- SUNTER, A.B. (1986). Solutions to the problem of unequal probability sampling without replacement. *International Statistical Review*, 54, 33-50.
- YATES, F., and GRUNDY, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society*, B, 15, 235-261.



# A View on Statistical Disclosure Control for Microdata

A.G. de WAAL and L.C.R.J. WILLENBORG<sup>1</sup>

## ABSTRACT

Problems arising from statistical disclosure control, which aims to prevent that information about individual respondents is disclosed by users of data, have come to the fore rapidly in recent years. The main reason for this is the growing demand for detailed data provided by statistical offices caused by the still increasing use of computers. In former days tables with relatively little information were published. Nowadays the users of data demand much more detailed tables and, moreover, microdata to analyze by themselves. Because of this increase in information content statistical disclosure control has become much more difficult. In this paper the authors give their view on the problems which one encounters when trying to protect microdata against disclosure. This view is based on their experience with statistical disclosure control acquired at Statistics Netherlands.

**KEY WORDS:** Statistical disclosure control; Microdata; Uniqueness.

## 1. INTRODUCTION

Statistical disclosure control (SDC) is becoming increasingly important as a result of the growing demand for information provided by statistical offices. The information released by these statistical offices can be divided into two major parts: tabular data and microdata. Whereas tables have been released traditionally by statistical offices, microdata sets are released only since fairly recently. In the past the users of data usually did not have the tools to analyze these microdata sets properly themselves. Nowadays every serious researcher is in possession of a powerful personal computer. Analyzing microdata is therefore no longer a privilege of the statistical office. The users of data can and want to analyze these microdata themselves. This creates non-trivial SDC-problems.

A key problem in the theory of SDC for microdata is the determination of the probability that a record in a released microdata set is re-identified. In order to estimate this probability a number of different approaches have been attempted. The aim of these attempts differ considerably. In some publications the aim was to gain a qualitative insight into the probability of re-identification of an unspecified record from a microdata set. In other publications the aim was set much higher, namely to obtain the probability that a specific record is re-identified. These are, of course, extreme cases. The former case is comparatively easy to solve, although still difficult. The latter case is more difficult and may be impossible to solve.

In this paper we give an overview of the problems for which Statistics Netherlands has attempted to provide a solution and problems of which the suggested solution has attracted our attention. We consider the problems and their outline of the solutions, while technical points are

skipped. The choice of the problems and the possible solutions we consider is heavily influenced by the experiences of Statistics Netherlands in the field of SDC.

The rest of this paper is organized as follows. Basic concepts are defined in Section 2. Preliminaries on SDC for microdata are the subject of Section 3. Our basic philosophy of SDC for microdata is discussed in Section 4. In Section 5 we describe the ideal situation for microdata: in this case we would have a probability for each record that this specific record can be re-identified. A somewhat less ideal situation is described in Section 6: in this case we have a probability for a data set that an unspecified record can be re-identified. In Section 7 we have to face reality: at the moment we do not have a good disclosure risk model and we have to be satisfied with heuristic arguments. In Section 8 we summarize our conclusions and suggest some possibilities for future research.

## 2. BASIC CONCEPTS

In this section a number of basic concepts are defined. We will assume that the statistical office wants to release a microdata set containing records of a sample of the population. Each record contains information about an individual entity. Such an entity could be a person, a household or a business enterprise. In the rest of this paper we will usually consider the individual entity to be a person, although this is not essential.

The two most important concepts in the field of SDC are re-identification and disclosure. Re-identification is said to occur if an attacker establishes a one-to-one relationship between a microdata record and a target individual with a sufficient degree of confidence. Following

<sup>1</sup> A.G. de Waal and L.C.R.J. Willenborg, Statistics Netherlands, Division of Research and Development, Department of Statistical Methods, P.O. Box 4000, 2270 JM Voorburg, The Netherlands (E-mail: TWAL@CBS.NL and LWLG@CBS.NL).

Skinner (1992) we distinguish between two kinds of disclosure. Re-identification disclosure occurs if the attacker is able to deduce the value of a sensitive variable for the target individual after this individual has been re-identified. Prediction disclosure (or attribute disclosure) occurs if the microdata enable the attacker to predict the value of a sensitive variable for some target individual with a sufficient degree of confidence. For prediction disclosure it is not necessary that re-identification has taken place. Most research so far has concentrated on re-identification disclosure. In this paper we will use the term disclosure to indicate re-identification disclosure unless stated otherwise.

Now, let us define what is meant by an identifying variable. A variable is called identifying if it can serve, alone or in combination with other variables, to re-identify some respondents by some user of the data. Examples of identifying variables are residence, sex, nationality, age, occupation and education. A subset of the set of identifying variables is the set of direct (or formal) identifiers. Examples of direct identifiers are name, address and public identification numbers. Direct identifiers must have been removed from a microdata set before it is released for else re-identification is very easy. Other identifiers in most cases do not have to be removed from the microdata set. A combination of identifying variables is called a key. The identifying variables that together constitute a key are also called key variables. A key value is a combination of scores on the identifying variables that together constitute the key.

In practice, determining whether or not a variable is identifying is a problem that can only be solved by sound judgment. No limitative list of intrinsically identifying variables exists, nor, for that matter, an unambiguous and well-defined set of rules to determine such variables. Selecting a set of identifying variables, and therefore of keys, is generally based on subjective assumptions about the population. Statistics Netherlands applies some criteria, like the visibility of the categories of a variable, to determine whether or not a variable is identifying, but these criteria do not provide a definite answer to this problem for all variables. Whether or not a variable is considered identifying is essentially a matter of judgment. In the remainder of this paper we will assume however that a set of keys has been determined.

The counterparts of identifying variables are the sensitive (or confidential) variables. A variable is called sensitive (or confidential) if some of the values represent characteristics a respondent would not like to be revealed about him. In principle, Statistics Netherlands considers all variables sensitive, but in practice some variables are considered more sensitive than others. Like in the case of identifying variables, determining whether or not a variable is sensitive can be solved only by sound judgment in practice. The variables sexual behavior and criminal past are generally considered sensitive, but for other variables this may depend on, for instance, cultural background. Keller and

Bethlehem (1992) give as an example the variable income. In the Netherlands income is considered sensitive, whereas in Sweden it is not. Moreover, there are variables which should be considered both identifying and sensitive. An example of such a variable is ethnic membership. However, in the literature it is usually assumed that the identifying and sensitive variables can be divided into disjoint sets. In the remainder of this paper we will also assume that a set of sensitive variables has been determined which is disjoint from the set of identifying variables.

By using information about the identifying variables a potential attacker can try to disclose information about sensitive variables. Note that this way of disclosure is only possible in case the link between the values of the identifying variables and the values of the sensitive variables has not been perturbed by noise in the data or by a technique like data-swapping.

To end this section, we give a definition of SDC. Statistical disclosure control aims to reduce the risk that sensitive information of individual persons can be disclosed to an acceptable level. What is acceptable depends on the policy of the data releaser. In order to reduce the risk of disclosure an estimate for the risk of disclosure would be very helpful although it is not a necessary requisite (*cf.* Section 7). Some research has been devoted to defining and estimating this risk of disclosure.

### 3. PRELIMINARIES ON SDC FOR MICRODATA

As a customer of a statistical office, the user of a microdata set should be satisfied with its quality. The user is usually not interested in individual records, but only in statistical results which can be drawn from the total set of records. For instance, he wants to examine tables he has produced himself from the microdata set.

Because a microdata set is meant for statistical analysis it is not necessary that each record in the set is correct. The statistical office has the possibility to perturb records, *e.g.*, by adding noise or by swapping parts of records between different records, in order to reduce the risk of re-identification. By perturbing records the risk of re-identification is reduced because even when a correct re-identification takes place the information which is disclosed may be incorrect. In any case the attacker cannot be sure that the disclosed information is correct. The statistical office 'only' has to guarantee that the statistical quality of, for instance, the tables the user wants to examine is high enough. This may be quite complicated to achieve in practice, however.

Although data perturbation methods may prove to be useful, for the time being Statistics Netherlands does not use them. To protect its microdata sets Statistics Netherlands applies local suppression and global recoding only.



When local suppression is applied some values of variables in some records are set to 'missing', *i.e.*, deleted from the microdata set. When global recoding is applied some variables are given a coarser categorization. In a first step, we try to protect a microdata set by means of global recoding. However, when protecting a microdata set entirely by means of global recodings would result in a considerable information loss, we apply local suppressions as well. In this way we try to avoid that too much information will be lost. It should be clear that local suppressions are only applied parsimoniously.

An advantage of local suppression and global recoding is that these techniques preserve the integrity of the data. A disadvantage of local suppression is that it introduces a bias, because extreme values will be locally suppressed. However, when local suppressions are only applied parsimoniously, this bias will be small.

From the SDC point of view a user of the data should also be looked upon as a potential attacker. Hence, it is useful to consider the ways in which disclosure can take place. An attacker tries to match records from the microdata set with records from an identification file or with individuals from his circle of acquaintances. An identification file is a file containing records with values on direct identifiers and values on some other identifiers of the microdata set. The latter identifiers may be used to match records from the released microdata set with records from the identification file. After matching the direct identifiers in the identification file can be used to determine whose record has been matched, and the sensitive variables in the released microdata set can be used to disclose information about this person. A circle of acquaintances is the set of persons in the population for which the attacker knows the values on a certain key from the microdata set. So, a circle of acquaintances could actually be an identification file, and vice versa. In the rest of this paper we will therefore use the terms 'identification file' and 'circle of acquaintances' interchangeably.

In order for re-identification of a record of an individual to occur the following conditions have to be satisfied:

- $C_1$ . The individual is unique on a particular key value  $K$ .
- $C_2$ . The individual belongs to an identification file or a circle of acquaintances of the attacker.
- $C_3$ . The individual is an element of the sample.
- $C_4$ . The attacker knows that the record is unique in the population on the key  $K$ .
- $C_5$ . The attacker comes across the record in the microdata set.
- $C_6$ . The attacker recognizes the record of the individual.

Whenever one of the conditions  $C_1$  to  $C_6$  does not hold, re-identification cannot be accomplished with absolute certainty. If either condition  $C_1$  or  $C_4$  does not hold, then a matching can be made but the attacker cannot be sure that this leads to a correct re-identification.

It is clear from the conditions  $C_1$  to  $C_6$  that a 'good' model for the risk of re-identification should incorporate aspects of both the data set and the user. When a Dutch microdata set is used by someone in, say, China who is essentially unfamiliar with the Dutch population, then the risk of re-identification is negligible. In order to re-identify someone in a microdata set it is necessary to acquire sufficient knowledge about the population. The amount of work that should be done to acquire this knowledge is proportional to the safety of the microdata set.

#### 4. A PHILOSOPHY OF SDC

It seems likely that the attention of a potential attacker is drawn by combinations of identifying variables that are rare in the sample or in the population. Combinations that occur quite often are less likely to trigger his curiosity. If he tries to match records deliberately then he will probably try to do this for key values that occur only a few times. If the user does not try to match records deliberately, but he knows an acquaintance with a rare key value then a record with that particular key value may trigger him to consider the possibility that this record belongs to this acquaintance. Moreover, the probability of a correct match is higher in case the number of persons that score on the matching key value is smaller. Finally, it is also very likely that among the persons that score on a rare key value there are many uniques if the key is augmented with an additional variable. Records that score on such rare combinations of identifying variables are therefore more likely to be re-identified.

In particular key values which occur only once in the population, *i.e.*, uniques in the population, can lead to re-identification. In the past emphasis was placed almost exclusively on uniqueness. It should be noted, however, that uniqueness is neither sufficient nor necessary for re-identification. If a person is unique in the population on certain key variables, but nobody realizes this, then this person may never be re-identified. If on the other hand this person is not unique in the population, but there is only one other person in the population with the same key, then this other person is, in principle, able to re-identify him. Furthermore, suppose a person is not unique, but belongs to a small group of people. Suppose also that the attacker happens to know information about him which is not considered to be identifying by the statistical office, but which is contained in the released microdata set, then it is very well possible that he is unique on the key combined with the new information. So, it is possible that a person is re-identified although he is not unique on the keys of identifying variables in the population. Finally, prediction disclosure may occur. That is, if a person is not unique in the population, but belongs to a group of people with (almost) the same score on a particular sensitive variable,



then sensitive information can be disclosed about this individual without actual re-identification. Prediction disclosure is not discussed further in this paper. For more information on prediction disclosure we refer to Skinner (1992), US Department of Commerce (1978), Duncan and Lambert (1986), and Cox (1986).

SDC should concentrate on key values that are rare in the population. A probability that information from a particular respondent, whose data are included in a microdata set, is disclosed should reflect the 'rareness' of the key value of this respondent's record. A probability for the event that information from an arbitrary respondent is disclosed should reflect the 'overall rareness' of the records in the data set. If there are many records in a microdata set of which the key value is rare, then the probability of disclosure for this data set should be high. In the next sections we will examine some attempts to incorporate these ideas within a mathematical framework.

## 5. RE-IDENTIFICATION RISK PER RECORD

In an ideal world (as far as SDC is concerned) a releaser of microdata would be able to determine a risk of re-identification for each record, *i.e.*, a probability that the respondent of this record can be re-identified. Such a risk per record would enable us to adopt the following strategy. First, order the records according to their risk of re-identification with respect to a single key. Second, select a maximum risk the statistical office is willing to accept. Finally, modify all the records for which the risk of re-identification with respect to the key chosen is too high. Repeat this procedure for each key in case there are more keys.

Unfortunately, we do not live in such an ideal world at the moment. However, steps towards the ideal situation have been made by Paass and Wauschkuhn (1985), and Fuller (1993). In Paass and Wauschkuhn (1985) it is assumed that a potential attacker has both a microdata file, released by a statistical office, and an identification file at his disposal. Between both files there may be many data incompatibilities. These data incompatibilities may be caused by *e.g.*, coding errors, by different definitions of categories or by 'noise' in the data. By assuming a probability distribution for these data incompatibilities and a disclosure scenario Paass and Wauschkuhn develop a sophisticated model to estimate the probability that a specific record from the microdata file is re-identified. The type of distribution of the errors that caused the data incompatibilities was assumed to be known to the attacker. The variance of the errors was assumed unknown to him. A potential attacker had to estimate this variance, on the basis of the (assumed) knowledge of the statistical production process. The model of Paass and Wauschkuhn is essentially based on discriminant analysis and cluster analysis.

Paass and Wauschkuhn distinguish between six different scenarios. Each scenario corresponds to a special kind of attacker. The number of records in the identification file and the information content of the identification file depend on the chosen scenario. An example of such a scenario is the journalist scenario, where a journalist selects records with extreme attribute combinations in order to re-identify respondents with the aim of showing that the statistical office fails to secure the privacy of its respondents.

Paass and Wauschkuhn apply their method to match records from the identification file with records from the microdata file. If the probability that a specific record from the identification file belongs to a specific record from the microdata set is high enough, then these two records are matched. This probability is the probability of re-identification per record, conditional on a particular disclosure scenario.

Müller, Blien, Knoche, Wirth *et al.* (1991) and Blien, Wirth and Müller (1992) applied the method recommended in Paass and Wauschkuhn (1985) to real data. When compared to simple matching, *i.e.*, a record is considered re-identified by an attacker if he succeeds in finding a unique value set in the microdata file which is identical to a value set in the identification file, the method suggested by Paass and Wauschkuhn turned out to be not superior. Apparently, the number of correctly matched records when applying the method by Paass and Wauschkuhn was in disagreement with the probability of re-identification per record.

In the context of masking procedures, *i.e.*, procedures for microdata disclosure limitation by adding noise to the microdata, Fuller (1993) obtained an expression for the probability that a specific record in the released microdata set is the same as a specific target record from an identification file. That is, an expression for the re-identification probability per record is derived. To derive this expression several assumptions are made. It is assumed that the data, the noise and errors in the data are normally distributed. Moreover, it is assumed that the covariance matrices of both the noise and the errors in the data are known to an attacker. Finally, it is assumed that the data have been obtained by simple random sampling. These assumptions allow Fuller (1993) to derive his expression for the re-identification probability by means of probability theoretical considerations. Unfortunately, the approach by Fuller has not been tested on real data yet. Hence, it is hard judge the applicability of this approach. For a comment on the approach by Fuller see Willenborg (1993).

Paass and Wauschkuhn (1985), and Fuller (1993) are mainly interested in the effects of noise that has (unintentionally and intentionally, respectively) been added to the data on the disclosure risk. A weak point of their respective approaches is the, implicit, assumption that the key is a high-dimensional one. Assuming a high-dimensional key implies that (almost) everyone in the population is unique. The probability that a combination or key value occurs more



than once in the population is negligible. This makes the computation of the probability of re-identification per record considerably easier. On the other hand, in case of low-dimensional keys it is not unlikely that certain key values occur many times in the population. Therefore, deriving a probability of re-identification per record for low-dimensional keys is much harder than for high-dimensional keys, because for high-dimensional keys the probability of statistical twins in the population is almost zero.

A good model for the re-identification risk per record does not appear to exist at the moment. In Section 6 we therefore consider less ambitious models, namely models for the re-identification risk per file.

## 6. RE-IDENTIFICATION RISK PER FILE

In a somewhat less ideal world a releaser of microdata would not be able to determine the risk of re-identification for each record, but he would be able to determine the risk that an unspecified record from the microdata set is re-identified. In this case, the statistical office should decide on the maximal risk it is willing to take when releasing a microdata set. If the actual risk is less than the maximal risk, then the microdata set can be released. If the actual risk is higher than the maximal risk, then the microdata set has to be modified. Determining which records have to be modified remains a problem, however.

A basic model to determine the probability that an arbitrary record from a microdata set is re-identified has been proposed by Mokken, Pannekoek and Willenborg (1989) and Mokken, Kooiman, Pannekoek and Willenborg (1992). In Mokken *et al.* (1989) only the case where there is a single researcher, an unstratified population and a single key is considered. It has been extended to include the cases of subpopulations, multiple researchers and multiple keys (*cf.* Willenborg 1990a; Willenborg 1990b; Mokken *et al.* 1992). The model of Mokken *et al.* (1992) takes three probabilities into account. The first probability,  $f$ , is equal to the sampling fraction. In other words,  $f$ , is the probability that a randomly chosen person from the population has been selected in the sample. The second probability,  $f_a$ , is the probability that a specific researcher who has access to the microdata knows the values of a randomly chosen person from the population on a particular key. The third probability,  $f_u$ , is the probability that a randomly chosen person from the population is unique in the population on a particular key. Combining these three probabilities,  $f$ ,  $f_a$  and  $f_u$ , the probability that a record from a microdata set is re-identified can be evaluated.

For each sample element a number of variables is measured. The values obtained by these measurements (scores) are collected in records, one for each sample element. It is assumed that the variables in the key are either categorical variables or variables for which the measurements fall into a finite number of categories.

Together, the records constitute a data set  $S$  that will be made available to a researcher  $R$ . We recall that whenever we use the term disclosure in fact re-identification disclosure is meant. The model of Mokken *et al.* (1989, 1992) does not take prediction disclosure into account.

In terms of the Paass and Wauschkunn (1985) set-up  $f_a$  and  $f_u$  together reflect the *Informationsgehalt der Überschneidungsmerkmale*, *i.e.*, the information content of the matching values. The various scenarios they consider differ in terms of  $f_a$  and  $f_u$ . In particular,  $f_u$  is influenced by the number of variables and the information content of these variables, *i.e.*, their categorization, an attacker has at his disposal to re-identify a record. The parameter  $f_a$  is determined by the number of records that are contained in the information file.

With respect to researcher  $R$  and key  $K$  there is a circle of acquaintances  $A$ . Obviously,  $A$  and its size  $|A|$  will depend on the particular researcher  $R$  as well as on the key  $K$  and the variables as registered and coded in the data set.

It is assumed that if conditions  $C_1$ ,  $C_2$  and  $C_3$  of the conditions for re-identification given in Section 3 hold, then conditions  $C_4$ ,  $C_5$  and  $C_6$  hold too. Condition  $C_4$  is a rather exacting one, but it can be introduced as an assumption for the sake of convenience in formulating a disclosure risk model. Note that it then yields a worst-case situation, in the sense that fallible perception and memory or other sources of ignorance, confusion and uncertainty for a potential discloser are excluded. Taken as an assumption together with  $C_5$  and  $C_6$  the implication is that the occurrence of any unique acquaintance  $E$  of  $R$  in data set  $S$  is equivalent to re-identification by  $R$ . It is assumed that re-identification of a record implies disclosure of confidential information. Thus re-identification can be treated as equivalent to disclosure. Implicitly, it is assumed that the link between the identifying variables and the sensitive variables has not been disturbed by a technique such as data-swapping.

Furthermore it is assumed that both the identifying and the confidential information are free of error or noise to researcher  $R$ , contrary to *e.g.*, Paass and Wauschkunn (1985), and Fuller (1993). Clearly, this assumption is unrealistic for most microdata sets.

The disclosure risk  $D_R$  for a certain microdata set  $S$  with respect to a certain researcher  $R$  and a certain key  $K$ , is defined to be the probability that the researcher makes at least one disclosure of a record in  $S$  on the basis of  $K$ . In order to apply a criterion based on the disclosure risk, the value of this quantity for a given data set has to be determined. An expression for this quantity can be derived on the basis of a set of assumptions.

In the model of Mokken *et al.* the following assumptions are made in addition to  $C_1 - C_6$ :

- $A_1$ . The circle of acquaintances  $A$  can be considered as a random sample from the population.
- $A_2$ . The data set  $S$  is a random sample from the population.



Assumption  $A_1$  serves to imply that the probability that a randomly chosen element from the population is an acquaintance of  $R$  is  $f_a = |A|/N$ , where  $N$  is the size of the population. As a consequence the expected number of unique elements in  $A$ ,  $|U_a|$ , is equal to  $f_a |U| = |A| f_u$ , where  $U$  is the set of unique persons in the population and  $|U|$  its size. Obviously assumption  $A_2$  implies that the probability that a specific unique element  $E$  is selected in the sample is  $f$ . These assumptions allow one to obtain a very simple expression for the disclosure risk  $D_R$  in terms of  $f$ ,  $f_a$  and  $f_u$ , namely

$$D_R = 1 - \exp(-Nff_af_u). \quad (1)$$

Two of the parameters in the model of Mokken *et al.* (1989, 1992),  $f_a$  and  $f_u$ , are unknown. The parameter  $f_a$  can be 'guestimated', *i.e.*, obtained by inspired guesswork, by assuming different scenarios an attacker may follow. A number of such scenarios has been described in Paass and Wauschkunn (1985) and Paass (1988). Evaluating  $f_a$  seems difficult, however. In order to estimate the other parameter,  $f_u$ , a number of models has been proposed in the literature. Models to estimate the number of uniques in the population, and hence  $f_u$ , that have been proposed include the Poisson-gamma model (Bethlehem, Keller and Pannekoek 1989; Mokken *et al.* 1989; Willenborg, Mokken and Pannekoek 1990; De Jonge 1990), the negative binomial superpopulation model (Skinner, Marsh, Openshaw and Wymer 1990), the Poisson-lognormal model (Skinner and Holmes 1992; Hoogland 1994), models based on equivalence classes (Greenberg and Zayatz 1992) and models based on modified negative binomial-gamma functions (Crescenzi 1992; Coccia 1992). As we have remarked in Section 4 not only the number of population uniques is important, but the numbers of cells with two, three, *etc.* persons are important as well. The Poisson-gamma model, the Poisson-lognormal model and the negative binomial superpopulation model can be applied to estimate the number of cells with two, three, *etc.* persons as well. It seems that the other models mentioned above can be extended in order to estimate these numbers. A major drawback is that the results are not very reliable in many cases.

From the model by Mokken *et al.* (1989, 1992) it is clear that the statistical office that disseminates the data is able to influence the risk of re-identification. The statistical office basically has two ways to do this. First of all, the size of the data set can be reduced, *i.e.*, the sampling fraction  $f$  can be reduced. A reduction of  $f$  implies a reduction of the risk. However, lowering  $f$  is generally undesirable, because usually  $f$  has to be reduced substantially to be effective. This implies that only a small part of the data available can be released. The second way in which the statistical office can influence the re-identification risk is by reducing the number of population uniques, *i.e.*, by reducing  $f_u$ . The fraction  $f_u$  depends on the information

provided by the key variables. The less information the key variables provide the less uniques there are in the population. In other words,  $f_u$  can be reduced by collapsing categories (global recoding) and by replacing values by missings (local suppression). Collapsing categories is a global action, because it generally affects many records; replacing values by missings is a local action because it affects only a few individual records. Usually, the loss in information when reducing  $f_u$  is considerably less than the loss in information when reducing  $f$ . Therefore, a statistical office will usually choose to control the re-identification risk by reducing  $f_u$  rather than reducing  $f$ . The third possibility of controlling the re-identification risk, *i.e.*, by reducing  $f_a$ , is not applied in practice, because  $f_a$  is difficult to model.

Although the model by Mokken *et al.* (1989, 1992) provides some insight in how to reduce the disclosure risk it can hardly be used as a basis for the protection of microdata sets. The reason for this is that the two parameters of the model,  $f_u$  and  $f_a$ , are often difficult to evaluate. Usually there is insufficient data available to estimate  $f_u$  and  $f_a$  accurately. We conclude that even a model for a re-identification risk for an entire microdata set is difficult to apply in practice. In Section 7 we therefore face reality in which we have no satisfactory model for either the re-identification risk per record or re-identification risk for an entire microdata set.

## 7. INTUITIVE RE-IDENTIFICATION RISK

In reality we are, unfortunately, forced to base SDC on heuristic arguments rather than on a solid theoretical basis. The SDC rules mentioned in this section all reduce the re-identification risk. It is, however, not possible to evaluate this reduction of the re-identification risk. At Statistics Netherlands, rules for SDC of microdata are based on testing whether scores on certain keys occur frequently enough in the population. A few problems arising here are the determination of the keys that have to be examined, the way to estimate the number of persons in the population that score on a certain key, to make operational the meaning of the phrase 'frequently enough' by determining *e.g.*, (a) threshold value(s), and how to determine appropriate SDC-measures.

Statistics Netherlands distinguishes between two kinds of microdata sets. The first kind is a so-called public use file. A public use file can be obtained by everybody. The keys that have to be examined for a public use file are all combinations of two identifying variables. The number of identifying variables is limited, and certain identifying variables, such as place of residence are not included in a public use file. Moreover, sampling weights have to be examined before they can be included in a public use file, because there are many situations in which weights can give additional information (*cf.* De Waal and Willenborg 1995a).



For instance, when a certain subpopulation is oversampled then this subpopulation can be recognized by the low weights associated with its members in the sample. Weights may only be published when they do not provide additional information that can be used for disclosure purposes. In case sampling weights are not considered suited for publication SDC measures should be taken, such as subsampling the units with a low weight in order to get a subsample in which all units have approximately the same weight. Because the weights are approximately equal assuming that they are exactly equal would introduce only a small error. The second kind of microdata set is a so-called microdata set for research. A microdata set for research can only be obtained by well-respected (statistical) research offices. The information content of a microdata set for research is much higher than that of a public use file. The number of identifying variables is not limited and an identifying variable such as place of residence may be included in a microdata set for research. Because of the high information content of a microdata set for research, researchers have to sign a declaration stating that they will protect any information about an individual respondent that might be disclosed by them. The keys that have to be examined for a microdata set for research consist of three-way combinations of variables describing a region with variables describing the sex, ethnic group or nationality of a respondent with an ordinary identifying variable.

The rules Statistics Netherlands applies for SDC are based on the following idea: a key value, *i.e.*, a combination of scores on the identifying variables that together constitute the key, is considered safe for release if the frequency that this key value occurs in the population is more than a certain threshold value  $d_0$ . This value  $d_0$  was chosen after a careful and extensive search considering many different values and comparing the records which have to be modified for each value of  $d_0$ . The value that leads to the 'most likely' set of records which have to be modified has been chosen to be the value of  $d_0$ . Which records are considered to be the 'most likely' ones to be modified is a matter of personal judgment.

When applying one of the above rules we are generally posed with the problem that we do not know the number of times that a key value occurs in the population. We only have the sample available to us. The population frequency of a key value has to be estimated based upon the sample. For large regions it is possible to use an interval estimator to test whether or not a key value occurs often enough in a region. This interval estimator is based on the assumption that the number of times that a key value occurs in the population is Poisson distributed (*cf.* Pannekoek 1995). However, for relatively small regions the number of respondents is low, which causes the estimator to have a high variance which in turn causes a lot of records to be modified. To estimate the number of times that a key value occurs in a small region we therefore suggest to apply

a point estimator. We will now discuss some possibilities for such an estimator.

A simple point estimator for the number of times that a certain key value occurs in a region is the direct point estimator. The fraction of a key value in a region  $i$  is estimated by the sample frequency of this key value in region  $i$  divided by the number of respondents in region  $i$ . The population frequency is then estimated by this estimated fraction multiplied by the number of inhabitants in region  $i$ . When the number of respondents in region  $i$  is low, which is often the case, the direct estimator is unreliable. Another point estimator is based on the assumption that the persons who score on a certain key value are distributed homogeneously over the population. In this case the fraction of a key value in region  $i$  can be estimated by the fraction in the entire sample. The advantage of this, so-called, synthetic, estimator is that the variance is much smaller than the variance of the direct estimator. Unfortunately, the homogeneity assumption is usually not satisfied which causes the estimator to be biased. However, a combined estimator can be constructed with both an acceptable variance and an acceptable bias by using a convex combination of the direct estimator and the synthetic estimator. Such a combined estimator has been tested in Pannekoek and de Waal (1995).

Another practical problem that deserves attention is top-coding of extreme values of continuous (sensitive) variables. These extreme values may lead to re-identification because these values are rare in the population. At the moment Statistics Netherlands uses an interval estimator to test whether there is a sufficient number of individuals in the population who score on a 'comparable' value of the continuous variable (*cf.* Pannekoek 1992). If this is the case, then the extreme value may be published, otherwise the extreme value must be suppressed. In order to apply this method in practice it remains to specify what is meant by 'sufficient' and by 'comparable'.

Some important practical problems occur when determining which protection measures should be taken when a microdata set appears to be unsafe. In that case the original data set must be modified in such a way that the information loss due to SDC-measures is as low as possible while the resultant data set is considered safe. In De Waal and Willenborg (1994a) and De Waal and Willenborg (1995b) a model for determining the optimal local suppressions is presented. Determining the optimal global recodings is much more difficult. Comparing the information loss due to global recodings to the information loss to local suppressions is already a problem. In De Waal and Willenborg (1995c) this latter problem is solved by using the entropy.

Currently a general purpose software package for SDC of microdata is being developed at Statistics Netherlands (*cf.* De Jong 1992; De Waal and Willenborg 1994b; Van Gelderen 1995; Pieters and De Waal 1995; De Waal and



Pieters 1995). The package, ARGUS, should enable the statistical office to analyze the data and to carry out suitable protection measures. It will consist of two separate parts:  $\mu$ -ARGUS for SDC of microdata and  $\tau$ -ARGUS for SDC of tabular data. The structure of the package is such that it will be possible to specify different disclosure control rules. This implies that ARGUS will be suited for other statistical offices too. Moreover, it will be possible to incorporate changes in the rules fairly easily in the package.

## 8. CONCLUSIONS

There is one important conclusion one can draw from this paper: SDC still offers a lot of possibilities for future research, despite the considerable amount of research that has been carried out to date. The theory of SDC for microdata has a number of gaps. Among the technical problems that remain to be solved are the following. When we want to release data for small regions we need an acceptable estimator for the number of times that a key value occurs in these regions. Such an estimator is difficult to construct, although the preliminary results obtained at Statistics Netherlands seem encouraging. An important practical problem is the determination of appropriate global recodings and local suppressions. Yet another one is the determination of the number of uniques, or more generally the number of rare frequencies, in the population. Some of the models proposed in Section 6 appear to be acceptable, but can probably be improved upon. An alternative approach is to determine which elements in the sample are unique in the population. In Verboon (1994), and Verboon and Willenborg (1995) this approach is examined. An extension of the model by Mokken *et al.* (1989, 1992) to estimate the risk of re-identification of a file is yet another problem to be solved. This extension should take into account that measurement errors have been made and that population uniqueness is not necessary in order to disclose information. Finally, a model to estimate the re-identification risk per record would be very welcome. In fact, it would yield a sound criterion to judge the safety of a microdata set. This criterion can guide one in producing safe microdata sets by applying SDC-measures such as global recoding and local suppression.

Apart from technical problems there are also some policy problems. Based on the policy that a statistical office wants to pursue the following decisions should be made. The combinations of variables that should be examined should be specified. Suitable threshold values should be selected.

More and better software must be developed in order to deal with time-consuming calculations. For microdata, software must be developed to indicate which records and variables must be modified, and how they should be modified, when applying a particular disclosure rule. At

the time of writing an international project on SDC is about to start. The participating institutions in this project are the Eindhoven University of Technology, the University of Manchester, the University of Leeds, the Office of Population Censuses and Surveys (OPCS), the Istituto Nazionale di Statistica (ISTAT), the Consorzio Padova Ricerche (CPR), and Statistics Netherlands. One of the major aims of the project is to develop software for the SDC of both microdata ( $\mu$ -ARGUS) and tabular data ( $\tau$ -ARGUS).

Finally, some very practical problems remain to be solved. An example of such a problem is the determination of a set of rules for selecting identifying variables. Such a set of rules would be a very valuable asset. Without these rules identifying variables are selected by making subjective choices. Developing such a set of rules is another goal of the above mentioned SDC-project.

## REFERENCES

- BETHLEHEM, J.A., KELLER, W.J., and PANNEKOEK, J. (1989). Disclosure control of microdata. *Journal of the American Statistical Association*, 85, 38-45.
- BLIEN, U., WIRTH, H., and MÜLLER, M. (1992). Disclosure risk for microdata stemming from official statistics. *Statistica Neerlandica*, 46, 69-82.
- COCCIA, G. (1992). Disclosure risk in Italian current population surveys. International Seminar on Statistical Confidentiality, Dublin.
- COX, L.H. (1986). Comment on Duncan and Lambert (1986). 19-21.
- CRESCENZI, F. (1992). Estimating population uniques; methodological proposals and applications on Italian census data. International Seminar on Statistical Confidentiality, Dublin.
- De JONG, W.A.M. (1992). ARGUS: An integrated system for data protection. International Seminar on Statistical Confidentiality, Dublin.
- De JONGE, G. (1990). The estimation of population unicity from microdata files (in Dutch), Internal note, Statistics Netherlands, Voorburg.
- De WAAL, A.G., and PIETERS, A.J. (1995). ARGUS user's guide. Report, Statistics Netherlands, Voorburg.
- De WAAL, A.G., and WILLENBORG, L.C.R.J. (1994a). Minimizing the number of local suppressions in a microdata set. Report, Statistics Netherlands, Voorburg.
- De WAAL, A.G., and WILLENBORG, L.C.R.J. (1994b). Development of ARGUS: past, present, future. Report, Statistics Netherlands, Voorburg.
- De WAAL, A.G., and WILLENBORG, L.C.R.J. (1995a). Statistical disclosure control and sampling weights. Report, Statistics Netherlands, Voorburg.
- De WAAL, A.G., and WILLENBORG, L.C.R.J. (1995b). Local suppression in statistical disclosure control and data editing. Report, Statistics Netherlands, Voorburg.

- De WAAL, A.G., and WILLENBORG, L.C.R.J. (1995c). Optimal global recoding and local suppression. Report, Statistics Netherlands, Voorburg.
- DUNCAN, G.T., and LAMBERT, D. (1986). Disclosure-limited data dissemination. *Journal of the American Statistical Association*, 81, 10-28.
- FULLER, W.A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics*, 9, 383-406.
- GREENBERG, B.V., and ZAYATZ, L.V. (1992). Strategies for measuring risk in public use microdata files. *Statistica Neerlandica*, 46, 33-48.
- HOOGLAND, J. (1994). Protecting microdata sets against statistical disclosure by means of compound Poisson distributions (in Dutch). Report, Statistics Netherlands, Voorburg.
- KELLER, W.J., and BETHLEHEM, J.A. (1992). Disclosure protection of microdata: problems and solutions. *Statistica Neerlandica*, 46, 5-19.
- MOKKEN, R.J., PANNEKOEK, J., and WILLENBORG, L.C.R.J. (1989). Microdata and disclosure risks, CBS Select 5, Statistical Essays, Staatsuitgeverij (The Hague), 181-200.
- MOKKEN, R.J., KOOIMAN, P., PANNEKOEK, J., and WILLENBORG, L.C.R.J. (1992). Disclosure risks for microdata. *Statistica Neerlandica*, 46, 49-67.
- MÜLLER, W., BLIEN, U., KNOCH, P., WIRTH, H. *et al.* (1991). *The Factual Anonymity of Microdata* (in German). Stuttgart: Metzler-Poeschel Verlag.
- PAASS G., and WAUSCHKUHN, U. (1985). Data access, data protection and anonymization – analysis potential and identifiability of anonymized individual data (in German). Gesellschaft für Mathematik und Datenverarbeitung, Oldenbourg-Verlag, Munich.
- PAASS, G. (1988). Disclosure risk and disclosure avoidance for microdata. *Journal of Business and Economic Studies*, 6, 487-500.
- PANNEKOEK, J. (1992). Disclosure control of extreme values of continuous identifiers (in Dutch). Report, Statistics Netherlands, Voorburg.
- PANNEKOEK, J. (1995). Statistical methods for some simple disclosure limitation rules. Report, Statistics Netherlands, Voorburg.
- PANNEKOEK, J., and de WAAL, A.G. (1995). Synthetic and combined estimators in statistical disclosure control. Report, Statistics Netherlands, Voorburg.
- PIETERS, A.J., and De WAAL, A.G. (1995). A demonstration of ARGUS. Report, Statistics Netherlands, Voorburg.
- SKINNER, S., MARSH, C., OPENSHAW, S., and WYMER, C. (1990). Disclosure avoidance for census microdata in Great Britain. *Proceedings of the 1990 Annual Research Conference*, U.S. Bureau of the Census, Washington, DC, 131-143.
- SKINNER, C.J. (1992). On identification disclosure and prediction disclosure for microdata. *Statistica Neerlandica*, 46, 21-32.
- SKINNER, C.J., and HOLMES, D.J. (1992). Modelling population uniqueness. International Seminar on Statistical Confidentiality, Dublin.
- US DEPARTMENT OF COMMERCE (1978). Report on statistical disclosure and disclosure avoidance techniques. Statistical Policy Working Paper 2, Washington DC.
- Van GELDEREN, R. (1995). ARGUS: Statistical disclosure control of survey data. Report, Statistics Netherlands, Voorburg.
- VERBOON, P. (1994). Some ideas for a masking measure for statistical disclosure control. Report, Statistics Netherlands, Voorburg.
- VERBOON, P., and WILLENBORG, L.C.R.J. (1995). Comparing two methods for recovering population uniques in a sample. Report, Statistics Netherlands, Voorburg.
- WILLENBORG, L.C.R.J. (1990a). Remarks on disclosure control of microdata. Report, Statistics Netherlands, Voorburg.
- WILLENBORG, L.C.R.J. (1990b). Disclosure risks for microdata sets: stratified populations and multiple investigators. Report, Statistics Netherlands, Voorburg.
- WILLENBORG, L.C.R.J. (1993). Discussion statistical disclosure limitation. *Journal of Official Statistics*, 9, 469-474.
- WILLENBORG, L.C.R.J., MOKKEN, R.J., and PANNEKOEK, J. (1990). Microdata and disclosure risks. *Proceedings of the 1990 Annual Research Conference*, U.S. Bureau of the Census, Washington DC, 167-180.





## CONTENTS

## TABLE DES MATIÈRES

## Volume 24, No. 1, March/mars 1996

Mikelis BICKIS, Susana BLEUER and Daniel KREWSKI	
On the estimation of the proportion of positives in a sequence of screening experiments .....	1
Hugh CHIPMAN	
Bayesian variable selection with related predictors .....	17
Lawrence JOSEPH, Alain C. VANDAL and David B. WOOLFSON	
Estimation in the multi-path change-point problem .....	37
André Robert DABROWSKI and Herold DEHLING	
Estimating conditional occupation time distributions for dependent sequences .....	55
Douglas P. WIENS	
Robust sequential designs for approximately linear models .....	67
Norbert HENZE	
Empirical distribution function goodness of fit tests for discrete models .....	81
Denna B. HAUNSPERGER	
Paradoxes in nonparametric tests .....	95
Peter Yi-Shi SHAO and William E. STRAWDERMAN	
Improving on truncated linear estimates of exponential and gamma scale variates .....	105
Samuel BAYOMOG, Xavier GUYON, Cécile HARDOUIN and Jianfeng YAO	
Test de différence de contrastes et somme pondérée de khideux .....	115
Yoav BENJAMIN and Abba M. KRIEGER	
Concepts and measures for skewness with data analytic implications .....	131

## Volume 24, No. 2, June/juin 1996

N. REID	
Likelihood and higher order approximations to tail areas: a review and annotated bibliography .....	141
Jiahua CHEN and J.D. KALBFLEISCH	
Penalized minimum distance estimates in finite mixture models .....	167
Brajendra C. SUTRADHAR and J.N.K. RAO	
Estimation of regression parameters in generalized linear models for clustered correlated data with measurement error .....	177
Bing LI and Ruben H. ZAMAR	
M-estimates of regression when scale is unknown and the error distribution is possibly asymmetric: a minimax results .....	193
R. B. ARELLANO-VALLE, H. BOLFARINE and F. VILCA-LABRA	
Ultrastructural elliptical models .....	207
R.L. DYKSTRA, Hammou EL BARMÍ, James M. GUFFEY and F.T. WRIGHT	
Nonhomogeneous Poisson processes as overhaul models .....	217
Edit GOMBAY	
The weighted sequential likelihood ratio .....	229
Feng-Gin SUN, Jean-Yves LARAMÉE and John S. RAMBERG	
On Spiring's normal loss function .....	241
Lee J. BAIN and Gaoxiong GAN	
Conditional maxima and inferences for the truncated exponential distribution .....	251
Y. YANG, A.K. GUPTA and T.T. NGUYEN	
Characterization theorems for some discrete distributions based on conditional structure .....	257
K. VIRASWAMI and N. REID	
Higher order asymptotics under model misspecification .....	263



# JOURNAL OF OFFICIAL STATISTICS

An International Quarterly Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey Methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

## Contents

### Volume 11, Number 4, 1995

Comparison Between Maximum Likelihood and Bayes Methods for Estimation of Binominal Probability with Sample Compositing <i>Yogendra P. Chaubey and Weiming Li</i> .....	379
Editing Statistical Records by Neural Networks <i>Svein Nordbotten</i> .....	391
Computer-assisted Personal Interviewing: An Experimental Evaluation of Data Quality and Costs <i>Reginald P. Baker, Norman M. Bradburn, and Robert A. Johnson</i> .....	413
Questionnaire Effects on Measurements of Race and Spanish Origin <i>Nancy Bates, Elizabeth A. Martin, Theresa J. DeMaio, and Manuel de la Puente</i> .....	433
A Content Analysis of Advance Letters from Expenditure Surveys of Seven Countries <i>Martin Luppès</i> .....	461
In Other Journals .....	481
Special Note .....	483
Editorial Collaborators .....	485
Index to Volume 11, 1995 .....	491

Lars Lyberg, U/LEDN, Statistics Sweden, S-115 81 Stockholm, Sweden







## GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue (Vol. 19, No. 1 and onward) of *Survey Methodology* as a guide and note particularly the following points:

### 1. Layout

- 1.1 Manuscripts should be typed on white bond paper of standard size ( $8\frac{1}{2} \times 11$  inch), one side only, entirely double spaced with margins of at least  $1\frac{1}{2}$  inches on all sides.
- 1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
- 1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

### 2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

### 3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as “exp(·)” and “log(·)”, etc.
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (e.g., w,  $\omega$ ; o, O; 0; 1, 1).
- 3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

### 4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
- 4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

### 5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.



## DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, prière d'examiner un numéro récent de *Techniques d'enquête* (à partir du vol. 19, n° 1) et de noter les points suivants:

1. **Présentation**
  - 1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8½ par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.
  - 1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.
  - 1.3 Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
  - 1.4 Les remerciements doivent paraître à la fin du texte.
  - 1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.

2. **Résumé**

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3. **Rédaction**
  - 3.1 Éviter les notes au bas des pages, les abréviations et les sigles.
  - 3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp(-) et log(-) etc.
  - 3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
  - 3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.
  - 3.5 Distinguer clairement les caractères ambigus (comme w, ω; o, O; 1, I).
  - 3.6 Les caractères italiques sont utilisés pour faire ressortir des mots. Indiquer ce qui doit être imprimé en italique en le soulignant dans le texte.

4. **Figures et tableaux**
  - 4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
  - 4.2 Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois.)

5. **Bibliographie**
  - 5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence.  
Exemple: Cochran (1977, p. 164).
  - 5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.







# JOURNAL OF OFFICIAL STATISTICS

An International Quarterly Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey Methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

## Contents

### Volume 11, Number 4, 1995

Comparison Between Maximum Likelihood and Bayes Methods for Estimation of Binominal Probability with Sample Compositing  
*Yogendra P. Chaubey and Weiming Li* ..... 379

Editing Statistical Records by Neural Networks  
*Svein Nordbotten* ..... 391

Computer-assisted Personal Interviewing: An Experimental Evaluation of Data Quality and Costs  
*Reginald P. Baker, Norman M. Bradburn, and Robert A. Johnson* ..... 413

Questionnaire Effects on Measurements of Race and Spanish Origin  
*Nancy Bates, Elizabeth A. Martin, Theresa J. DeMaio, and Manuel de la Puente* ..... 433

A Content Analysis of Advance Letters from Expenditure Surveys of Seven Countries  
*Martin Luppes* ..... 461

In Other Journals ..... 481

Special Note ..... 483

Editorial Collaborators ..... 485

Index to Volume 11, 1995 ..... 491

Lars Lyberg, U/LEDN, Statistics Sweden, S-115 81 Stockholm, Sweden

## TABLE DES MATIÈRES

## CONTENTS

1	Mikelis BICKIS, Susana BLEUER et Daniel KREWSKI On the estimation of the proportion of positives in a sequence of screening experiments .....
17	Hugh CHIPMAN Bayesian variable selection with related predictors .....
37	Lawrence JOSEPH, Alain C. VANDAL et David B. WOOLFSON Estimation in the multi-path change-point problem .....
55	André Robert DABROWSKI et Herold DEHLING Estimating conditional occupation time distributions for dependent sequences .....
67	Douglas P. WIENS Robust sequential designs for approximately linear models .....
81	Norbert HENZE Empirical distribution function goodness of fit tests for discrete models .....
95	Denna B. HAUENSPERGER Paradoxes in nonparametric tests .....
105	Peter Yi-Shi SHAO et William E. STRAWDERMAN Improving on truncated linear estimates of exponential and gamma scale variates .....
115	Samuel BAYOMOG, Xavier GUYON, Cécile HARDOUIN et Jianfeng YAO Test de différence de contrastes et somme pondérée de k hideux .....
131	Yoav BENJAMIN et Abba M. KRIEGER Concepts and measures for skewness with data analytic implications .....
<b>Volume 24, No. 1, March/mars 1996</b>	
141	N. REID Likelihood and higher order approximations to tail areas: a review and annotated bibliography .....
167	Jiahua CHEN et J.D. KALBFLEISCH Penalized minimum distance estimates in finite mixture models .....
177	Brajendra C. SUTRADHAR et J.N.K. RAO Estimation of regression parameters in generalized linear models for clustered correlated data with measurement error .....
193	Bing Li et Ruben H. ZAMAR M-estimates of regression when scale is unknown and the error distribution is possibly asymmetric: a minimax results .....
207	R. B. ARELLANO-VALLE, H. BOLFARINE et F. VILCA-LABRA Ultrastuctural elliptical models .....
217	R. L. DYKSTRA, Hammou EL BARM, James M. GUFFEY et F.T. WRIGHT Nonhomogeneous Poisson processes as overhaul models .....
229	Edit GOMBAY The weighted sequential likelihood ratio .....
241	Feng-Gin SUN, Jean-Yves LARAMEE et John S. RAMBERG On Spiriting's normal loss function .....
251	Lee J. BAIN et Gaoxiong GAN Conditional maxima and inferences for the truncated exponential distribution .....
257	Y. YANG, A.K. GUPTA et T.T. NGUYEN Characterization theorems for some discrete distributions based on conditional structure .....
263	K. VIRASWAMI et N. REID Higher order asymptotics under model misspecification .....

- US DEPARTMENT OF COMMERCE (1978). Report on statistical disclosure and disclosure avoidance techniques. Statistical Policy Working Paper 2, Washington DC.
- Van GELDEREN, R. (1995). ARGUS: Statistical disclosure control of survey data. Rapport, Statistics Netherlands, Voorburg.
- VERBOON, P. (1994). Some ideas for a masking measure for statistical disclosure control. Rapport, Statistics Netherlands, Voorburg.
- VERBOON, P., et WILLENBORG, L.C.R.J. (1995). Comparing two methods for recovering population uniques in a sample. Rapport, Statistics Netherlands, Voorburg.
- WILLENBORG, L.C.R.J. (1993). Discussion statistical disclosure sets: stratified populations and multiple investigators. Rapport, Statistics Netherlands, Voorburg.
- WILLENBORG, L.C.R.J. (1990b). Disclosure risks for microdata sets: stratified populations and multiple investigators. Rapport, Statistics Netherlands, Voorburg.
- WILLENBORG, L.C.R.J. (1990a). Remarks on disclosure control of microdata. Rapport, Statistics Netherlands, Voorburg.
- WILLENBORG, L.C.R.J., MOKKEN, R.J., et PANNENKOEK, J. (1990). Microdata and disclosure risks. *Proceedings of the 1990 Annual Research Conference*, U.S. Bureau of the Census, Washington DC, 167-180.



article, un projet international sur le CSD, auquel participent l'Eindhoven University of Technology, l'University of Manchester, l'University of Leeds, l'Office of Population Censuses and Surveys (OPCS), l'Istituto Nazionale di Statistica (ISTAT), le Consorzio Padova Ricerche (CPR) et Statistics Netherlands, est sur le point de débiter. Un des objectifs principaux du projet consiste à mettre au point un logiciel applicable au CSD des microdonnées ( $\mu$ -ARGUS) et des données tabulaires ( $\tau$ -ARGUS). Enfin, des problèmes d'un ordre très pratique doivent être résolus, comme la détermination d'un ensemble de règles de sélection des variables identifiantes. Cet ensemble de règles constituerait un atout précieux. En effet, sans ces règles, les variables identifiantes sont choisies de façon subjective. L'élaboration d'un tel ensemble de règles est un autre objectif du projet de CSD susmentionné.

## BIBLIOGRAPHIE

- BETHLEHEM, J.A., KELLER, W.J., et PANNEKOEK, J. (1989). Disclosure control of microdata. *Journal of the American Statistical Association*, 85, 38-45.
- BLIEN, U., WIRTH, H., et MÜLLER, M. (1992). Disclosure risk for microdata stemming from official statistics. *Statistica Neerlandica*, 46, 69-82.
- COCCIA, G. (1992). Disclosure risk in Italian current population surveys. International Seminar on Statistical Confidentiality, Dublin.
- COX, L.H. (1986). Commentaries sur Duncan et Lambert (1986). 19-21.
- CRESCENZI, F. (1992). Estimating population uniques; methodological proposals and applications on Italian census data. International Seminar on Statistical Confidentiality, Dublin.
- De JONG, W.A.M. (1992). ARGUS: An integrated system for data protection. International Seminar on Statistical Confidentiality, Dublin.
- De WAAL, A.G., et PIETERS, A.J. (1995). ARGUS user's guide. Rapport, Statistics Netherlands, Voorburg.
- De WAAL, A.G., et WILLENBORG, L.C.R.J. (1994b). Minimizing the number of local suppressions in a microdata set. Rapport, Statistics Netherlands, Voorburg.
- De WAAL, A.G., et WILLENBORG, L.C.R.J. (1994a). Statistical disclosure control and sampling weights. Rapport, Statistics Netherlands, Voorburg.
- De WAAL, A.G., et WILLENBORG, L.C.R.J. (1995b). Local suppression in statistical disclosure control and data editing. Rapport, Statistics Netherlands, Voorburg.
- De WAAL, A.G., et WILLENBORG, L.C.R.J. (1995c). Optimal global recoding and local suppression. Rapport, Statistics Netherlands, Voorburg.
- DUNCAN, G.T., et LAMBERT, D. (1986). Disclosure-limited data dissemination. *Journal of the American Statistical Association*, 81, 10-28.
- FULLER, W.A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics*, 9, 383-406.
- GREENBERG, B.V., et ZAVAYATZ, L.V. (1992). Strategies for measuring risk in public use microdata files. *Statistica Neerlandica*, 46, 33-48.
- HOOGLAND, J. (1994). Protecting microdata sets against statistical disclosure by means of compound Poisson distributions (en néerlandais). Rapport, Statistics Netherlands, Voorburg.
- KELLER, W.J., et BETHLEHEM, J.A. (1992). Disclosure protection of microdata: problems and solutions. *Statistica Neerlandica*, 46, 5-19.
- MOKKEN, R.J., PANNEKOEK, J., et WILLENBORG, L.C.R.J. (1989). Microdata and disclosure risks, CBS Select 5, Statistical Essays, Staatsuitgeverij (La Haye), 181-200.
- MOKKEN, R.J., KOOMAN, P., PANNEKOEK, J., et WILLENBORG, L.C.R.J. (1992). Disclosure risks for microdata. *Statistica Neerlandica*, 46, 49-67.
- MÜLLER, W., BLIEN, U., KNOCH, P., WIRTH, H., et coll. (1991). *The Factual Anonymity of Microdata* (en allemand). Stuttgart: Metzler-Poeschel Verlag.
- PAASS G., et WAUSCHKUH, U. (1985). Data access, data protection and anonymization – analysis potential and identifiability of anonymized individual data (en allemand). Gesellschaft für Mathematik und Datenverarbeitung, Oldenbourg-Verlag, Munich.
- PAASS, G. (1988). Disclosure risk and disclosure avoidance for microdata. *Journal of Business and Economic Studies*, 6, 487-500.
- PANNEKOEK, J. (1992). Disclosure control of extreme values of continuous identifiers (en néerlandais). Rapport, Statistics Netherlands, Voorburg.
- PANNEKOEK, J. (1995). Statistical methods for some simple disclosure limitation rules. Rapport, Statistics Netherlands, Voorburg.
- PANNEKOEK, J., et de WAAL, A.G. (1995). Synthetic and combined estimators in statistical disclosure control. Rapport, Statistics Netherlands, Voorburg.
- PIETERS, A.J., et De WAAL, A.G. (1995). A demonstration of ARGUS. Rapport, Statistics Netherlands, Voorburg.
- SKINNER, S., MARSH, C., OPENSHAW, S., et WYMER, C. (1990). Disclosure avoidance for census microdata in Great Britain. *Proceedings of the 1990 Annual Research Conference*, U.S. Bureau of the Census, Washington, DC, 131-143.
- SKINNER, C.J. (1992). On identification disclosure and prediction disclosure for microdata. *Statistica Neerlandica*, 46, 21-32.
- SKINNER, C.J., et HOLMES, D.J. (1992). Modelling population uniqueness. International Seminar on Statistical Confidentiality, Dublin.

l'échantillon de cette valeur clé dans la région  $i$  par le nombre de répondants dans la région  $i$ . Puis, on estime sa fréquence dans la population en multipliant la fraction estimée par le nombre d'habitants dans la région  $i$ . Quand le nombre de répondants dans la région  $i$  est faible, ce qui est souvent le cas, l'estimateur direct n'est pas fiable. Un autre estimateur ponctuel se fonde sur l'hypothèse selon laquelle les personnes qui produisent une valeur clé particulière sont réparties uniformément dans la population. Dans ce cas, on peut estimer la fraction de la valeur clé dans la région  $i$  en déterminant la fraction dans l'échantillon complet. L'avantage que présente cet estimateur dit «synthétique» tient au fait que sa variance est beaucoup plus faible que celle de l'estimateur direct. Malheureusement, l'hypothèse quant à l'homogénéité de la population n'étant généralement pas satisfaite, l'estimateur est biaisé. Cependant, on peut créer un estimateur combiné, dont tant la variance que le biais sont acceptables, par combinaison convexe de l'estimateur direct et de l'estimateur synthétique. Pannekoeck et de Waal (1995) ont testé un estimateur combiné de ce type.

Le camouflé des valeurs extrêmes des variables continues (délicates) est un autre problème pratique qui mérite d'être examiné. Ces valeurs extrêmes peuvent mener à la réidentification, car elles sont rares au sein de la population. Pour le moment, à Statistics Netherlands, nous nous servons d'un estimateur d'intervalle pour vérifier si la population contient un nombre suffisant de personnes pour lesquelles on observe une valeur «comparable» de la variable continue (consulter Pannekoeck, 1992), mais il est possible que nous appliquions un estimateur ponctuel à l'avenir. Si le nombre est suffisant, la valeur extrême peut être publiée. Sinon, elle doit être supprimée. Afin de mettre la méthode en pratique, il reste à préciser ce qu'on entend par «suffisant» et par «comparable».

D'importants problèmes d'ordre pratique surviennent lors de la détermination des mesures à prendre pour protéger un ensemble de microdonnées qui semble poser un risque. Le cas échéant, il convient de modifier l'ensemble de données original, de façon à réduire au minimum la perte d'information causée par les mesures de CSD tout en produisant un ensemble de données jugées sans risque. Un modèle permettant de déterminer les suppressions locales optimales est présenté par De Waal et Willenborg (1994a), et De Waal et Willenborg (1995b). La détermination du recodage global optimal est, par contre, beaucoup plus difficile. Comparer la perte d'information due au recodage global à celle qu'entraîne les suppressions locales pose déjà, en soi, un problème. De Waal et Willenborg (1995c) le résolvent en s'appuyant sur le concept d'entropie.

Statistics Netherlands est en train de mettre au point un progiciel d'usage général pour le CSD relatif aux microdonnées (consulter De Jong, 1992; De Waal et Willenborg, 1994b; Van Gelderen, 1995; Pieters et De Waal, 1995; De Waal et Pieters, 1995). Le progiciel, baptisé ARGUS, devrait permettre au bureau de la statistique d'analyser les données et d'exécuter les mesures de protection pertinentes. Il comprendra deux éléments distincts:  $\mu$ -ARGUS,

pour le CSD des microdonnées et  $\tau$ -ARGUS, pour le CSD des données tabulaires. Le progiciel est conçu de façon à pouvoir préciser diverses règles de contrôle de la divulgation. Donc, ARGUS pourra être utilisé par d'autres bureaux de la statistique. Qui plus est, il sera possible de modifier assez facilement les règles incluses dans le progiciel.

### 8. CONCLUSIONS

Une des premières conclusions qui se dégagent du précédent article est la suivante: malgré le nombre d'études déjà obtenus à Statistics Netherlands paraissent encourageants. En pratique, déterminer quels sont les recodages globaux et les suppressions locales appropriés constituent un problème important. La détermination du nombre d'éléments uniques ou, de façon plus générale, du nombre de fréquences rares dans la population en est un autre. Certains modèles proposés à la section 6 restent acceptables, mais pourraient sans doute être améliorés. Une autre méthode consiste à déterminer quels éléments de l'échantillon sont uniques dans la population. Verboon (1994) et Verboon et Willenborg (1995) ont examiné cette méthode. L'extension du modèle proposé par Mokken et coll. (1989, 1992) pour estimer le risque de réidentification d'un fichier est un autre problème non résolu. L'extension du modèle devrait tenir compte de ce que des erreurs de mesure ont été commises et que l'unicité de la population n'est pas toujours une condition nécessaire pour qu'il y ait lieu la divulgation d'information. Enfin, un modèle permettant d'estimer le risque de réidentification par enregistrement serait le bienvenu. En fait, un tel modèle fournirait un critère valable pour juger de la sécurité d'un ensemble de microdonnées. Ce critère servirait de guide pour produire des ensembles de microdonnées sûrs, grâce à l'application de mesures de CSD telles que le recodage global et la suppression locale.

Outre les problèmes techniques, des problèmes de politique restent à résoudre. Compte tenu de la politique qu'il se propose de suivre, le bureau de la statistique doit préciser quelles combinaisons de variables devraient être examinées et choisir des valeurs limites appropriées.

De nouveaux logiciels, plus performants, doivent être mis au point pour venir à bout plus facilement des calculs très longs. En ce qui concerne les microdonnées, il convient de mettre au point des logiciels qui, en appliquant une règle de divulgation particulière, indiquent quels enregistrements et quelles variables doivent être modifiés, et de quelle manière. Au moment où nous rédigeons le présent



près le même poids. Puisque les poids sont presque égaux, émettre l'hypothèse qu'ils le sont parfaitement n'entraîne qu'une faible erreur. Le deuxième type d'ensemble de microdonnées est l'ensemble de microdonnées dit « de recherche ». Un ensemble de microdonnées de recherche ne peut être obtenu que par des bureaux de recherche (statistique) dont la réputation est bien établie. Le contenu informatif d'un tel ensemble est beaucoup plus important que celui d'un fichier public. Le nombre de variables identifiées n'y est pas limité et une variable identifiée catrice telle que le lieu de résidence peut y être incluse. En raison du contenu informatif détaillé de ces fichiers, les chercheurs doivent signer une déclaration précisant qu'ils s'engagent à protéger tout renseignement susceptible d'être divulgué sur des répondants particuliers. Les clés qu'il convient d'examiner pour un ensemble de microdonnées de recherche correspondent à des combinaisons triples de variables décrivant une région, de variables décrivant le sexe, le groupe ethnique ou la nationalité d'un répondant, et d'une variable identifiante ordinaire.

Les règles qu'applique Statistics Netherlands en matière de CSD se fondent sur le principe selon lequel la diffusion d'une valeur clé, c'est-à-dire une combinaison de valeurs des variables identifiantes qui, ensemble, constituent la clé, est jugée sans danger si la fréquence à laquelle cette valeur clé se manifeste dans la population dépasse un certain seuil  $d_0$ . Cette valeur  $d_0$  a été choisie après un examen minutieux et extensif, consistant à envisager un grand nombre de valeurs distinctes, et à comparer les enregistrements qui doivent être modifiés pour chaque valeur de  $d_0$ . La valeur qui produit l'ensemble « le plus probable » d'enregistrements qui doivent être modifiés a été choisie comme valeur de  $d_0$ . La sélection de l'ensemble « le plus probable » d'enregistrements à modifier est une question de jugement.

Quand nous appliquons les règles susmentionnées, nous nous heurtons généralement au problème de ne pas connaître le nombre de fois qu'une valeur clé survient dans la population. Nous ne disposons que de l'échantillon. La fréquence d'une valeur clé au sein de la population doit être estimée à partir de cet échantillon. Dans le cas de grandes régions, il est possible de se servir d'un estimateur d'intervalle pour vérifier si une valeur clé se manifeste suffisamment souvent dans une région donnée. Cet estimateur d'intervalle est défini en supposant que le nombre de fois qu'une valeur clé survient dans la population obéit à la loi de Poisson (consulter Pannekoek, 1995). Cependant, dans le cas des régions relativement petites, le nombre de répondants étant faible, la variance de l'estimateur est élevée, situation qui, à son tour, oblige à modifier un grand nombre d'enregistrements. Par conséquent, pour estimer le nombre de fois qu'une valeur clé se manifeste dans une petite région, nous proposons d'utiliser un estimateur ponctuel. Examinons maintenant certains de ces estimateurs.

L'estimateur ponctuel direct constitue un moyen simple d'estimer le nombre de fois qu'une valeur clé particulière est observée dans une région. On estime la fraction d'une valeur clé dans une région  $i$  en divisant la fréquence dans

plutôt que  $f$ . La troisième méthode de contrôle du risque de réidentification, c.-à-d. la réduction de  $f_n$ , n'est pas appliquée en pratique, car il est difficile de modéliser  $f_n$ . Si le modèle de Mokken et coll. (1989, 1992) donne une idée des mesures qui peuvent être prises pour réduire le risque de divulgation, il peut difficilement servir de fondement à la protection des ensembles de microdonnées, car les deux paramètres du modèle,  $f_n$  et  $f_a$ , sont souvent difficiles à évaluer. Habituellement, on ne dispose pas de suffisamment de données pour les estimer avec précision. Nous concluons que même un modèle visant à calculer le risque de réidentification pour un ensemble complet de microdonnées est difficile à appliquer en pratique. À la section 7, nous faisons face à la réalité, à savoir le fait que nous ne disposons d'aucun modèle satisfaisant pour calculer le risque de réidentification par enregistrement ou pour un ensemble complet de microdonnées.

## 7. CONTRÔLE INTUITIF DU RISQUE DE RÉIDENTIFICATION

En réalité, nous sommes obligés de fonder le CSD sur des arguments heuristiques plutôt que sur des concepts théoriques solides. Les règles de CSD mentionnées dans la présente section permettent toutes de réduire le risque de réidentification. Néanmoins, il n'est pas possible d'évaluer la réduction. À Statistics Netherlands, les règles de CSD relatives aux microdonnées s'appuient sur des tests visant à déterminer si la fréquence de certaines valeurs clés est suffisamment élevée au sein de la population. Les problèmes qui se posent ici consistent à déterminer quelles clés il faut examiner et comment il faut estimer le nombre de membres de la population qui produisent une valeur clé particulière, à rendre opérationnelle la phrase « fréquence suffisamment élevée » en déterminant, par exemple, une ou plusieurs valeurs seuils, et enfin, à décider quelles sont les mesures de CSD appropriées.

Statistics Netherlands distingue deux types d'ensembles de microdonnées. Le premier est un fichier dit d'usage public. N'importe qui peut obtenir un tel fichier. Les clés qu'il convient d'examiner dans le cas d'un fichier public correspondent à toutes les combinaisons possibles de variables identifiantes prises deux à deux. Le nombre de variables identifiantes est limité et certaines, comme le lieu de résidence, ne sont pas incluses dans un fichier public. Il faut en outre examiner les poids d'échantillon-nage avant de les inclure dans le fichier, car très souvent, ils peuvent fournir des renseignements supplémentaires (consulter De Waal et Willemborg, 1995a). Par exemple, une sous-population particulière qu'on a suréchantillonné risque d'être reconnue à cause du faible poids attribué à ses membres dans l'échantillon. Les poids ne peuvent être publiés que s'ils ne fournissent aucun renseignement supplémentaire susceptible de faciliter la divulgation. Si on estime que les poids ne peuvent être publiés, il convient de prendre des mesures de CSD, telles que le sous-échantillonnage des unités dont le poids est faible, afin d'obtenir un sous-échantillon où toutes les unités ont à peu



Deux des paramètres du modèle de Mokken et coll. (1989, 1992),  $f_a$  et  $f_n$ , sont inconnus. On peut «estimer subjectivement» le paramètre  $f_a$ , c.-à-d. obtenir sa valeur par divination éclairée, en postulant divers scénarios qu'un attaquant est susceptible de suivre. Plusieurs de ces scénarios ont été décrits par Paass et Wauschkuhn (1985) et par Paass (1988). Néanmoins, il paraît difficile d'évaluer  $f_a$ . En ce qui concerne l'estimation de l'autre paramètre,  $f_n$ , plusieurs modèles ont été proposés. Ceux qui ont été préconisés pour estimer le nombre d'éléments uniques dans la population, donc  $f_n$ , incluent la distribution gamma de Poisson (Behlhelm, Keller et Pannekoeck 1989; Mokken et coll., 1989; Willenborg, Mokken et Pannekoeck 1990; De Jonge, 1990), le modèle de superpopulation à distribution binomiale négative (Skinner et coll., 1990), la distribution normale logarithmique de Poisson (Skinner et Holmes, 1992; Hoogland, 1994), des modèles fondés sur les classes d'équivalence (Greenberg et Zayat, 1992) et des modèles fondés sur des fonctions gamma binomiales négatives modifiées (Crescenzi, 1992; Coccia, 1992). Comme nous l'avons mentionné à la section 4, non seulement le nombre de populations uniques est important, mais aussi celui de cellules comptant deux, trois, etc. personnes. Il semble que les autres modèles susmentionnés puissent être élargis afin d'estimer ces nombres. Un inconvénient important tient au fait que, dans de nombreux cas, les résultats ne sont pas très fiables.

Le modèle de Mokken et coll. (1989, 1992) montre clairement que le bureau de la statistique qui diffuse les données peut prendre des mesures pour modifier le risque de réidentification, et ce, essentiellement de deux façons. En premier lieu, il est possible de réduire la taille de l'ensemble de données, c.-à-d. la fraction de sondage  $f$ . Une diminution de  $f$  implique une diminution du risque. Cependant, il n'est généralement pas souhaitable de réduire  $f$ , car la réduction doit être considérable pour que la mesure soit efficace. Donc, le cas échéant, seule une faible proportion des données disponibles peuvent être diffusées. En second lieu, le bureau de la statistique peut modifier le risque de réidentification en réduisant le nombre d'éléments uniques dans la population, c.-à-d. en réduisant  $f_n$ . La fraction  $f_n$  dépend de l'information que fournissent les variables clés. La population compte d'autant moins d'éléments uniques que les variables clés fournissent moins de renseignements. Autrement dit, on peut réduire  $f_n$  en regroupant les catégories (recodage global) et en remplaçant certaines valeurs par l'indication «manquante» (suppression locale). Le regroupement des catégories est une mesure globale, car elle touche généralement de nombreux enregistrements; le remplacement de certaines valeurs par une lacune est une mesure locale, car elle n'en concerne que quelques-uns. Ordinairement, la perte d'information est nettement moins importante quand on réduit  $f_n$  que quand on réduit  $f$ . Par conséquent, le bureau de la statistique décide généralement de contrôler le risque de réidentification en réduisant  $f_n$ .

Les auteurs présument que, si les conditions  $C_1$ ,  $C_2$  et  $C_3$  de la liste de conditions de réidentification énumérées à la section 3 sont satisfaites, les conditions  $C_4$ ,  $C_5$  et  $C_6$  sont remplies également. La condition  $C_4$  est relativement astreignante, mais, par souci de commodité, peut-être introduite à titre d'hypothèse quand on formule un modèle du risque de divulgation. Il convient de noter qu'un tel modèle représente le pire scénario, en ce sens que les défauts de perception ou de mémoire, ou d'autres sources d'ignorance, de confusion et d'incertitude liées à un divulgateur éventuel sont exclues. Quand on prend simultanément pour hypothèse les conditions  $C_5$  et  $C_6$ , on signifie que l'apparition de toute connaissance unique  $E$  de  $R$  dans l'ensemble de données  $S$  équivaut à la réidentification par  $R$ . Les auteurs supposent que la réidentification d'un enregistreur sous-entend la divulgation de renseignements confidentiels. Donc, ils peuvent considérer que la réidentification est équivalente à la divulgation. Implicitement, ils supposent que le lien entre les variables identifiantes et les variables délicates n'a pas été perturbé par une méthode telle que l'échange de données.

Qui plus est, contrairement à Paass et Wauschkuhn (1985), et à Fuller (1993), ils supposent que le chercheur  $R$  dispose de données tant identifiantes que confidentielles exemptes d'erreur ou de bruit. Manifestement, cette hypothèse est irréaliste pour la plupart des ensembles de microdonnées.

Le risque de divulgation  $D_R$  lié à un ensemble partiel de microdonnées  $S$  en ce qui a trait à un chercheur particulier  $R$  et à une clé particulière  $K$  se définit comme la probabilité que le chercheur divulgue au moins un des enregistrements de  $S$  en se basant sur  $K$ . Pour pouvoir appliquer un critère fondé sur le risque de divulgation, il est nécessaire de déterminer cette grandeur pour un ensemble de données particulier. En se basant sur un ensemble d'hypothèses, il est possible d'établir une équation permettant de calculer cette grandeur.

En plus des hypothèses  $C_1$  à  $C_6$  mentionnées précédemment, le modèle de Mokken et coll. se fonde sur les hypothèses suivantes:

- A1. Le cercle de connaissances  $A$  peut être considéré comme un échantillon aléatoire de la population.
- A2. L'ensemble de données  $S$  est un échantillon aléatoire de la population.

L'hypothèse A1 sous-entend que l'expression de la probabilité qu'un élément choisi au hasard au sein de la population soit une connaissance de  $R$  est  $f_a = |A|/N$ , où  $N$  représente la taille de la population. Par conséquent, le nombre prévu d'éléments uniques dans  $A$ ,  $|U_a|$ , est égal à  $f_a |U| = |A|/f_n$ , où  $U$  représente l'ensemble de personnes uniques au sein de la population et  $|U|$ , sa taille. De toute évidence, l'hypothèse A2 implique que la probabilité qu'un élément unique  $E$  particulier soit choisi dans l'échantillon est égal à  $f$ . Ces hypothèses permettent d'obtenir une expression très simple du risque  $D_R$  en fonction de  $f$ ,  $f_a$  et  $f_n$ , à savoir

$$D_R = 1 - \exp(-Nff_af_n). \quad (1)$$



Dans un monde un peu moins parfait, un diffuseur de microdonnées ne serait pas capable de déterminer le risque de réidentification pour chaque enregistré, mais pourrait calculer le risque qu'un enregistré indétectablement soit reconnu. Dans ces conditions, le bureau de la statistique doit décider du

## 6. RISQUE DE RÉIDENTIFICATION PAR FICHIER

En ce qui a trait aux méthodes de camouflages, c'est-à-dire les méthodes qui visent à limiter la divulgation en ajoutant un bruit aux microdonnées, Fuller (1993) propose une expression qui permet de calculer la probabilité qu'un enregistré appartenant à l'ensemble de microdonnées diffusé soit le même qu'un enregistré ciblé d'un fichier d'identification, autrement dit la probabilité de réidentification par enregistré. Il émet plusieurs hypothèses pour arriver à cette équation. En premier lieu, il suppose que la loi de distribution des données, du bruit et des erreurs est normale. En outre, il présume que l'attaquant connaît les matrices de covariances du bruit et des erreurs qui affectent les données. Enfin, il suppose que les données ont été obtenues par échantillonnage aléatoire simple. Ces hypothèses permettent à Fuller (1993) d'obtenir l'équation utilisée pour calculer la probabilité de réidentification grâce à des considérations théoriques au sujet des probabilités. Malheureusement, sa méthode n'a pas encore été éprouvée au moyen de données réelles. Donc, il est difficile de porter un jugement sur son applicabilité. Le lecteur trouvera dans Willenborg (1993) un commentaire de la méthode préconisée par Fuller.

Paas et Wauschkun (1985) et Fuller (1993) s'intéressent surtout aux effets qu'a sur le risque de divulgation le bruit qui est ajouté (involontairement et volontairement, respectivement) aux données. Un des points faibles de leurs méthodes respectives tient à l'hypothèse, implicite, que le nombre de dimensions de la clé est élevé. En effet, supposer que le nombre de dimensions de la clé est élevé sous-entend que (pratiquement) tous les membres de la population sont uniques. La probabilité qu'une combinaison ou une valeur clé se manifeste plus d'une fois au sein de la population est pratiquement nulle, situation qui rend le calcul de la probabilité de réidentification par enregistré beaucoup plus facile. En revanche, si le nombre de dimensions des clés est faible, il n'est pas improbable que certaines valeurs clés se manifestent fréquemment dans la population. Calculer la probabilité de réidentification par enregistré dans ces conditions est beaucoup plus difficile que quand le nombre de dimensions des clés est élevé et que la probabilité qu'il existe des paires statistiques au sein de la population est virtuellement nulle.

Les modèles du risque de réidentification par enregistré-tremment proposés jusqu'à présent ne semblant pas satisfaisants, à la section 6, nous envisageons des modèles moins ambitieux, qui permettent de calculer le risque de réidentification par fichier.

Un risque maximal qu'il est disposé à courir quand il diffuse un ensemble de microdonnées. Ce dernier peut être diffusé si le risque réel est inférieur au risque maximal, mais doit être modifié si le risque dépasse la limite. Toutefois, déterminer quels enregistrés doivent être modifiés est un exercice qui demeure problématique.

Mokken, Pannekoeck et Willenborg (1989) et Mokken, Kooiman, Pannekoeck et Willenborg (1992) ont proposé un modèle de base pour calculer la probabilité qu'un enregistré appartenant à un ensemble de microdonnées soit reconnu. Mokken et coll. (1989) envisagent uniquement le scénario comprenant un seul chercheur, une population non stratifiée et une clé unique. Ce modèle a été élargi afin d'inclure les cas de sous-populations, de chercheurs multiples et de clés multiples (consulter Willenborg, 1990a; Willenborg, 1990b; Mokken et coll., 1992). Le modèle de Mokken et coll. (1992) tient compte de trois probabilités. La première,  $f$ , est égale à la fraction de sondage. Autrement dit,  $f$  représente la probabilité qu'une personne choisie au hasard dans la population soit sélectionnée dans l'échantillon. La deuxième,  $f_a$ , représente la probabilité qu'un chercheur particulier ayant accès aux microdonnées connaisse les valeurs qui correspondent à une personne choisie au hasard au sein de la population pour une clé particulière. La troisième,  $f_n$ , est la probabilité qu'une personne choisie au hasard au sein de la population soit unique dans cette population en regard d'une clé particulière. En combinant ces trois probabilités,  $f$ ,  $f_a$  et  $f_n$ , on peut calculer la probabilité qu'un des enregistrés d'un ensemble de microdonnées soit réidentifié. Plusieurs variables sont mesurées pour chaque élément de l'échantillon. Les valeurs obtenues pour ces diverses mesures sont regroupées en enregistrés correspondant chacun à un élément particulier de l'échantillon. Mokken et ses collaborateurs supposent que les variables incluses dans la clé sont soit des variables nominales, soit des variables dont les mesures se classent dans un nombre fini de catégories. Regroupés, les enregistrés constituent un ensemble de données  $S$  qui est mis à la disposition d'un chercheur  $R$ . Rappelons que, chaque fois que nous utilisons le terme divulgation, nous entendons en fait divulgation par réidentification. Le modèle de Mokken et coll. (1989, 1992) ne tient pas compte de la divulgation par prédiction. Aux termes du modèle de Paas et Wauschkun (1985), les probabilités  $f_a$  et  $f_n$  réunies reflètent l'*Informationsgehalt der Überschneidungsmerkmale*, c.-à-d. le contenu informatif des valeurs apparées. Les divers scénarios qu'ils envisagent se distinguent en ce qui a trait à  $f_a$  et à  $f_n$ . Plus précisément, la valeur de  $f_n$  dépend du nombre de variables qu'un attaquant a à sa disposition pour réidentifier un enregistré, ainsi que du contenu informatif, c.-à-d. la valeur du paramètre  $f_a$  est fonction du nombre d'enregistrements que contient le fichier d'information.

En ce qui concerne le chercheur  $R$  et la clé  $K$ , il existe un cercle de connaissances  $A$ . Manifestement,  $A$  et sa dimension  $|A|$  dépendent du chercheur  $R$  et de la clé  $K$ , ainsi que des variables enregistrées et codées dans l'ensemble de données.



du risque qu'ils soient réidentifiés en regard d'une clé unique. En second lieu, déterminer le risque maximum que le bureau de la statistique est prêt à accepter. Enfin, modifier tous les enregistrements pour lesquels le risque de réidentification en regard de la clé choisie est trop élevé. Répéter la procédure pour chaque clé, s'il en existe plus d'une.

Malheureusement, à l'heure actuelle, ce monde idéal nous échappe. Cependant, Paass et Wauschkuhn (1985), et Fuller (1993) nous en rapprochent de quelque pas. Paass et Wauschkuhn (1985) supposent qu'un attaquant éventuel a à sa disposition un fichier de microdonnées diffusé par un bureau de la statistique, ainsi qu'un fichier d'identification. Bon nombre des données de ces deux fichiers peuvent être incompatibles. Les incompatibilités peuvent être le résultat d'erreurs de codage, de définitions différentes des catégories ou d'un «bruit» dans les données. Paass et Wauschkuhn émettent une hypothèse quant à la loi de distribution de ces données incompatibles et au mode de divulgation, et établissent un modèle complexe pour estimer la probabilité qu'un enregistreur particulier du fichier de microdonnées soit reconnu. Ils présument que l'attaquant connaît la loi de distribution des erreurs qui ont causé l'incompatibilité des données. Ils supposent en outre qu'il ne connaît pas la variance des erreurs. Un attaquant éventuel doit estimer cette variance, en se fondant sur sa connaissance (tenue pour acquise) de la méthode de production des données statistiques. Le modèle de Paass et Wauschkuhn repose essentiellement sur l'analyse descriptive et sur l'analyse typologique.

Paass et Wauschkuhn distinguent six scénarios qui correspondent chacun à un type particulier d'attaque. Le nombre d'enregistrements dans le fichier d'identification et le contenu informatif de ce fichier dépendent du scénario choisi. Un exemple de scénario est celui du journaliste qui sélectionne des enregistrements où figurent des combinaisons d'attributs extrêmes pour réidentifier des répondants, dans le but de prouver que le bureau de la statistique ne réussit pas à sauvegarder la vie privée de ses répondants.

Paass et Wauschkuhn appliquent leur méthode en vue d'apparier des enregistrements du fichier d'identification à des enregistrements du fichier de microdonnées. Quand la probabilité qu'un enregistreur particulier du fichier d'identification correspond à un enregistreur particulier du fichier de microdonnées est suffisamment forte, les deux enregistrements sont apparés. Cette probabilité représente la probabilité de réidentification par enregistrement, sous réserve d'un scénario de divulgation particulier.

Müller, Blien, Knoche, Wirth et coll. (1991) et Blien, Wirth et Müller (1992) ont appliqué la méthode préconisée par Paass et Wauschkuhn (1985) à des données réelles. Cette méthode ne s'est pas avérée meilleure que le simple apparierement, en vertu duquel un enregistrement est considéré réidentifié par un attaquant si ce dernier réussit à déceler dans le fichier de microdonnées un ensemble de valeurs uniques identique à un ensemble de valeurs figurant dans le fichier d'identification. Apparemment, le nombre d'enregistrements apparés correctement en appliquant la méthode de Paass et Wauschkuhn ne concordait pas avec la probabilité de réidentification par enregistrement.



d'enregistrer entre plusieurs de ceux-ci, par exemple, afin de diminuer le risque de réidentification. La perturbation des enregistrements diminue le risque, car, même si une réidentification a effectivement lieu, l'information divulguée n'est pas nécessairement correcte. En tout cas, l'attaquant ne peut être sûr que l'information divulguée est correcte. Le bureau de la statistique doit garantir «uniquement» que la qualité statistique des tableaux que l'utilisateur souhaite examiner, pour s'en tenir à l'exemple choisi, est suffisante. Néanmoins, en pratique, cela peut s'avérer difficile.

Bien que les méthodes de perturbation des données soient parfois utiles, Statistics Netherlands ne les utilise pas à l'heure actuelle. Pour protéger les ensembles de microdonnées, nous ne recourons qu'à la suppression locale et au recodage global. Dans le cas de la suppression locale, certaines valeurs des variables de certains enregistrements sont déclarées «manquantes», c.-à-d. supprimées de l'ensemble de microdonnées. Dans celui du recodage global, certaines variables sont catégorisées plus grossièrement. Nous essayons d'abord de protéger l'ensemble de microdonnées par recodage global. Cependant, si le recours à cette seule méthode entraîne une perte d'information considérable, nous effectuons également des suppressions locales afin d'éviter que la perte soit excessive. Nous tenons néanmoins à préciser que les suppressions locales sont effectuées partiellement.

L'avantage de la suppression locale et du recodage global tient à ce que ces méthodes préservent l'intégrité des données. La suppression locale a toutefois pour inconvénient d'introduire un biais, car les valeurs extrêmes sont supprimées localement. Cependant, si les suppressions locales sont effectuées partiellement, le biais est faible.

Du point de vue du CSD, l'utilisateur des données devrait aussi être considéré comme un attaquant éventuel. Par conséquent, il est utile d'imaginer de quelles façons la divulgation peut avoir lieu. Un attaquant essaye généralement d'apparier les enregistrements de l'ensemble de microdonnées à ceux d'un fichier d'identification ou à des personnes appartenant à son cercle de connaissances. Un fichier d'identification est un fichier dont les enregistrements contiennent des valeurs pour les identificateurs directs ainsi que pour d'autres identificateurs de l'ensemble de microdonnées. Les seconds identificateurs peuvent servir à apparier les enregistrements de l'ensemble de microdonnées diffusé à ceux du fichier d'identification. Après l'appariement, les identificateurs directs qui figurent dans le fichier d'identification peuvent être utilisés pour déterminer à quelle personne correspond l'enregistrement apparié, puis, les variables délicates de l'ensemble de microdonnées diffusé peuvent servir à divulguer l'information au sujet de la personne en question. Le cercle de connaissances de l'attaquant est l'ensemble des membres de la population pour lesquels il connaît les valeurs d'une clé particulière de l'ensemble de microdonnées. Donc, un cercle de connaissances pourrait effectivement jouer le rôle de fichier d'identification, et vice versa. Dans la suite de l'article, nous utiliserons indifféremment les expressions «fichier d'identification» et «cercle de connaissances».

#### 4.

### UNE CONCEPTION DU CONTRÔLE STATISTIQUE DE LA DIVULGATION

Pour que l'attaquant puisse reconnaître la personne à laquelle correspond un enregistrement particulier, il faut que les conditions ci-après soient satisfaites :

C<sub>1</sub>. La personne est unique en ce qui a trait à une valeur clé particulière K.

C<sub>2</sub>. La personne est incluse dans un fichier d'identification ou appartient à un cercle de connaissances de l'attaquant.

C<sub>3</sub>. La personne est un élément de l'échantillon.

C<sub>4</sub>. L'attaquant sait que, en ce qui concerne la clé K, l'enregistrement est unique au sein de la population.

C<sub>5</sub>. L'attaquant découvre l'enregistrement dans l'ensemble de microdonnées.

C<sub>6</sub>. L'attaquant reconnaît l'enregistrement de la personne.

Si une des conditions C<sub>1</sub> à C<sub>6</sub> n'est pas remplie, la réidentification ne peut être accomplie de façon absolument certaine. Si la condition C<sub>1</sub> ou C<sub>4</sub> n'est pas vérifiée, un appariement peut être effectué, mais l'attaquant ne peut être certain qu'il aboutira à une réidentification correcte. Il est manifeste, quand on examine la liste des conditions C<sub>1</sub> à C<sub>6</sub>, qu'un «bon» modèle du risque de réidentification doit englober certaines caractéristiques non seulement des données, mais aussi de l'utilisateur. Si un ensemble de microdonnées d'origine hollandaise est utilisé, disons, en Chine, par quelqu'un qui, fondamentalement, ne connaît pas la population hollandaise, le risque de réidentification est négligeable. Afin de reconnaître une personne sur laquelle des données figurent dans un ensemble de microdonnées, il est nécessaire de bien connaître la population. L'effort qui doit être déployé pour acquérir cette connaissance est proportionnelle à la sécurité de l'ensemble de microdonnées.

L'attention d'un attaquant éventuel est vraisemblablement attirée par des combinaisons rares de variables identifiantes dans l'échantillon ou dans la population. Les combinaisons relativement fréquentes sont moins susceptibles de susciter sa curiosité. S'il tente délibérément d'apparier des enregistrements, il s'efforcera probablement de le faire pour des valeurs clés peu fréquentes. Il se peut aussi qu'il n'essaie pas délibérément d'apparier des enregistrements, mais que, sachant qu'une de ses connaissances possède une valeur clé rare, la découvre d'un enregistrement contenant cette valeur clé le pousse à essayer de déterminer s'il se rapporte à la personne en question. En outre, la probabilité qu'un appariement soit correct est d'autant plus élevée que le nombre de personnes pour lesquelles on enregistre la valeur clé à apparier est faible. Enfin, il est également fort probable que, parmi les personnes associées à une valeur clé rare, beaucoup deviennent uniques quand on ajoute une autre variable à la clé. Les enregistrements qui contiennent de telles combinaisons rares de variables identifiantes sont donc plus susceptibles d'être reconnus.



peut être une personne, un ménage ou une entreprise. Dans la suite de l'exposé, nous considérerons, en règle générale, que l'entité particulière est une personne, bien que cela ne soit pas essentiel.

Dans le domaine du CSD, les deux concepts les plus importants sont la réidentification et la divulgation. On dit qu'il y a réidentification quand un attaquant établit, avec un degré de confiance suffisant, un lien univoque entre un enregistré de microdonnées et un individu cible. À l'exemple de Skinner (1992), nous distinguons deux types de divulgation. La divulgation par réidentification n'est pas nécessaire. Jusqu'à présent, la plupart des travaux de recherche ont mis l'accent sur la divulgation par réidentification. Dans le présent article, sauf avis contraire, c'est à cette dernière que fait allusion le terme divulgation.

Définissons maintenant ce que nous entendons par variable identifiante. Une variable est dite identifiante si, seule ou combinée à d'autres variables, elle permet à un utilisateur de données de reconnaître certains répondants. À titre d'exemple, mentionnons le lieu de résidence, le sexe, la nationalité, l'âge, la profession et le niveau de scolarité. Un sous-ensemble de l'ensemble des variables identifiantes est celui des identificateurs directs (ou formels), dont le nom, l'adresse et le numéro d'identité sont des exemples. Les identificateurs directs doivent être supprimés d'un ensemble de microdonnées avant la diffusion de ce dernier, sinon la réidentification serait très facile. Dans la plupart des cas, seule la suppression de ces indicateurs est nécessaire. Une combinaison de variables identifiantes est appelée une clé. Les variables identifiantes qui, ensemble, constituent une clé sont aussi appelées variables clés. Une valeur clé est une combinaison de valeurs enregistrées pour les variables identifiantes qui, ensemble, constituent la clé.

En pratique, déterminer si une variable est ou non identifiante est un problème qui ne peut être résolu qu'en faisant preuve de jugement. Il n'existe aucune liste limitative de variables identifiantes intrinsèques, ni, d'ailleurs, aucun ensemble non ambigu et bien défini de règles permettant de repérer de telles variables. La sélection d'un ensemble de variables identifiantes, donc, de clés, est généralement fondée sur des hypothèses subjectives au sujet de la population. Statistics Netherlands applique certains critères, comme la visibilité des catégories d'une variable, pour déterminer si celle-ci est identifiante, mais ces critères ne permettent pas de résoudre catégoriquement la question dans tous les cas. Qu'une variable soit ou ne soit pas considérée identifiante est essentiellement une question de jugement. Dans le reste de l'article, nous supposons qu'un ensemble de clés a été déterminé.

Les variables délicates (ou confidentielles) sont la contrepartie des variables identifiantes. Une variable est dite délicate (ou confidentielle) si certaines de ses valeurs représentent des caractéristiques qu'un répondant ne souhaiterait pas qu'on dévoile. En théorie, Statistics Netherlands considère que toutes les variables sont déli-cates, mais, en pratique, juge certaines plus délicates que d'autres. Comme dans le cas des variables identifiantes, déterminer si une variable est ou non délicate est une question qu'on ne peut résoudre, en pratique, qu'en faisant preuve de jugement. Ainsi, on considère, en règle générale, que le comportement sexuel et les antécédents criminels sont des variables délicates. Par contre, dans d'autres cas, la décision dépend, par exemple, du contexte culturel. À titre d'exemple, Keller et Bethlehem (1992) mentionnent la variable de revenu, jugée délicate aux Pays-Bas mais non en Suède. En outre, certaines variables, comme l'appar-tenance ethnique, devraient être considérées à la fois iden-tifiantes et délicates. Cependant, les auteurs qui traitent de ces questions émettent ordinairement l'hypothèse qu'il est possible de séparer les variables identifiantes et délicates pour former des ensembles disjointes. Dans la suite de l'exposé, nous supposons qu'on a déterminé un ensemble de variables délicates qui est disjoint de l'ensemble de variables identifiantes.

En se servant de renseignements sur les variables iden-tifiantes, un attaquant éventuel peut essayer de divulguer des renseignements sur des variables délicates. Il convient de noter que ce mode de divulgation n'est possible que si le lien entre les valeurs des variables identifiantes et des variables délicates n'a pas été perturbé par un bruit dans les données ou par une méthode telle que l'échange de données.

Pour conclure la présente section, définissons le contrôle statistique de la divulgation. Ce dernier a pour but de réduire à un niveau acceptable le risque que des renseigne-ments confidentiels sur certaines personnes soient divul-gués. La définition de l'acceptabilité est fonction de la politique du diffuseur de données. Afin de réduire le risque de divulgation, il serait utile de pouvoir estimer ce dernier, mais il ne s'agit pas d'une condition nécessaire (voir la section 7). Certains travaux de recherche ont été consacrés à la définition et à l'estimation du risque de divulgation.

### 3. PRÉLIMINAIRES DU CSD RELATIF AUX MICRODONNÉES

En tant que client d'un bureau de la statistique, l'utili-sateur d'un ensemble de microdonnées doit être satisfait de la qualité de ce dernier. Ordinairement, il s'intéresse non pas à des enregistrements particuliers, mais aux résul-tats statistiques qu'il peut tirer de l'ensemble, comme les tableaux qu'il produit lui-même à partir de ce dernier. Puisqu'un ensemble de microdonnées est destiné à l'analyse statistique, il n'est pas nécessaire que chaque enregistrement de l'ensemble soit correct. Le bureau de la statistique peut donc perturber les enregistrements, en ajoutant un bruit ou en échangeant certains éléments



## Apérçu du contrôle statistique de la divulgation des microdonnées

A.G. de WAAL et L.C.R.J. WILLENBORG

## RÉSUMÉ

Les problèmes que pose le contrôle statistique de la divulgation, lequel a pour but d'empêcher les utilisateurs des données de divulguer des renseignements sur des répondants particuliers, se sont multipliés rapidement au cours des dernières années. La situation est due principalement à l'augmentation de la demande de données détaillées provenant des bureaux de la statistique, elle-même causée par l'accroissement continu de l'usage des ordinateurs. Auparavant, ces bureaux produisaient des tableaux contenant relativement peu d'information. Aujourd'hui, par contre, les utilisateurs de données demandent des tableaux beaucoup plus détaillés et, qui plus est, des microdonnées à analyser eux-mêmes. Or, l'augmentation du contenu informatif des données rend le contrôle statistique de la divulgation beaucoup plus difficile. Les auteurs se fondent sur l'expérience qu'ils ont acquise dans le domaine du contrôle statistique de la divulgation à Statistics Netherlands pour exposer les problèmes qu'il faut, selon eux, surmonter quand on essaie de protéger les microdonnées contre la divulgation.

MOTS CLES: Contrôle statistique de la divulgation; microdonnées; unicité.

## 1. INTRODUCTION

Le contrôle statistique de la divulgation (CSD) est un domaine qui acquiert de plus en plus d'importance en raison de l'intérêt croissant pour les renseignements produits par les bureaux de la statistique. Ces renseignements se subdivisent en deux grandes catégories, à savoir les données tabulaires et les microdonnées. Si la diffusion de tableaux est une activité de longue date des bureaux de la statistique, celle des ensembles de microdonnées, quant à elle, est relativement récente. Par le passé, les utilisateurs de données ne possédaient généralement pas les outils nécessaires pour analyser correctement les ensembles de microdonnées. Aujourd'hui, par contre, tout chercheur qui se respecte possède un ordinateur personnel puissant. L'analyse des microdonnées n'est donc plus le privilège des bureaux de la statistique puisque les utilisateurs de données sont désormais capables de l'effectuer eux-mêmes et en manifestent l'intention. Les problèmes que pose cette situation en matière de CSD sont loin d'être bénins. Un des principaux obstacles auxquels se heurtent les

théoriciens du CSD des microdonnées consiste à calculer la probabilité qu'un enregistrement inclus dans un ensemble diffusé de microdonnées soit reconnu. Diverses méthodes, dont les objectifs varient considérablement, ont été proposées pour estimer cette probabilité. Certains chercheurs tentent simplement d'estimer qualitativement la probabilité qu'un enregistrement indéterminé d'un ensemble de microdonnées soit reconnu, tandis que d'autres, dont les visées sont beaucoup plus ambitieuses, cherchent à déterminer la probabilité qu'un enregistrement précis le soit. Il s'agit évidemment de deux cas extrêmes. Le premier problème, quoique d'une certaine complexité, est assez facile à résoudre. Le second, plus difficile, pourrait ne pas avoir de solution.

## 2. CONCEPTS DE BASE

Dans la présente section, nous commençons par définir plusieurs concepts de base. Nous supposons, pour ce faire, que le bureau de la statistique souhaite diffuser un ensemble de microdonnées renfermant des enregistrements relatifs à un échantillon de la population. Chaque enregistrement contient des renseignements sur une entité particulière, qui

La suite de l'exposé se présente comme il suit. À la section 2, nous définissons les concepts de base. À la section 3, nous décrivons les préliminaires du CSD des microdonnées. À la section 4, nous exposons notre théorie de base sur le CSD des microdonnées. À la section 5, nous envisageons la situation idéale, c'est-à-dire celle où on peut calculer la probabilité de réidentification de chaque enregistré. À la section 6, nous décrivons une situation un peu moins parfaite consistant à calculer la probabilité qu'un enregistré indetermine d'un ensemble de données soit reconnu. Puis, à la section 7, nous faisons face à la réalité, à l'existence, à l'heure actuelle, d'un bon modèle du risque de divulgation qui oblige à se satisfaire d'arguments heuristiques. Enfin, à la section 8, nous présentons nos conclusions ainsi que des questions qui mériteraient d'être étudiées plus en détail.

Dans le présent article, nous donnons un aperçu des problèmes qu'on s'est efforcé de résoudre à Statistics Netherlands et dont la solution proposée a retenu notre attention. Nous nous limitons à décrire les problèmes et leur solution, en laissant de côté les aspects purement techniques. Le choix des problèmes et des solutions que nous examinons est en grande partie fonction des expériences vécues à Statistics Netherlands en ce qui concerne

A.G. de Waal et L.C.R.J. Willenborg, Statistics Netherlands, Division of Research and Development, Department of Statistical Methods, C.P. Box 4000, 2270 JM Voorburg, Pays-Bas (courriel électronique: [TWAL@CBS.NL](mailto:TWAL@CBS.NL) et [LWL@CBS.NL](mailto:LWL@CBS.NL)).

Résultats des simulations, plan simple, stratification et stratification mobile – fin

$\rho^2$	Plan	Paramètres	$E_{sim} \widehat{Var} \hat{y}$	$Var \hat{y}$	$EQM_{sim}^{\hat{y}}$
0.0	sm	$M = 2.11N/n$	0.01319	0.01319	0.01328
	srs		0.01315	0.01316	0.01332
	strat	$H = 16$	0.01315	0.01308	0.01331
0.2	sm	$M = 2.11N/n$	0.01038	0.01036	0.01021
	srs		0.01334	0.01334	0.01025
	strat	$H = 16$	0.01034	0.01034	0.01025
0.4	sm	$M = 2.11N/n$	0.00796	0.00796	0.00792
	srs		0.01316	0.01316	0.01323
	strat	$H = 16$	0.00790	0.00801	0.00794
0.6	sm	$M = 2.11N/n$	0.00572	0.00573	0.00561
	srs		0.01315	0.01316	0.01299
	strat	$H = 16$	0.00568	0.00572	0.00563
0.8	sm	$M = 2.11N/n$	0.00295	0.00294	0.00290
	srs		0.01317	0.01316	0.01325
	strat	$H = 16$	0.00287	0.00288	0.00285
1.0	sm	$M = 2.11N/n$	0.00048	0.00048	0.00048
	srs		0.01317	0.01316	0.01335
	strat	$H = 16$	0.00037	0.00034	0.00034
0.0	sm	$M = 1.09N/n$	0.01325	0.01316	0.01310
	srs		0.01313	0.01316	0.01317
	strat	$H = 32$	0.01201	0.01239	0.01302
0.2	sm	$M = 1.09N/n$	0.01070	0.01062	0.01064
	srs		0.01313	0.01316	0.01316
	strat	$H = 32$	0.00972	0.01018	0.01083
0.4	sm	$M = 1.09N/n$	0.00807	0.00803	0.00811
	srs		0.01315	0.01316	0.01309
	strat	$H = 32$	0.00732	0.00751	0.00803
0.6	sm	$M = 1.09N/n$	0.00538	0.00534	0.00536
	srs		0.01315	0.01316	0.01310
	strat	$H = 32$	0.00484	0.00484	0.00543
0.8	sm	$M = 1.09N/n$	0.00283	0.00281	0.00276
	srs		0.01317	0.01316	0.01283
	strat	$H = 32$	0.00255	0.00276	0.00280
1.0	sm	$M = 1.09N/n$	0.00016	0.00016	0.00017
	srs		0.01317	0.01316	0.01304
	strat	$H = 32$	0.00012	0.00007	0.00011

BBBBINGTON, A.C. (1975). A simple method of drawing a sample without replacement. *Applied Statistics*, 24, 136.

COCHRAN, W.G. (1977). *Sampling Techniques*. New York: Wiley.

DEVILLE, J.-C., et GROSBRAS, J.-M. (1987). Algorithmes de tirage. Dans *Les sondages*. Droesbeke, J.-J., Fichet, B., et Tassi, P. (éds.). Paris: Economica, 209-233.

FAN, C.T., MULLER, M.E., et REZUCHA, I. (1962). Development of sampling plans by using sequential (item by item) selection techniques and digital computers. *Journal of the American Statistical Association*, 57, 387-402.

HÄJKE, J. (1971). Comment on an essay of D. Basu. Dans *Foundations of Statistical Inference*. Godambe V.P., et Sprott, D.A. (éds). Toronto: Holt, Rinehart et Winston.

MCLEOD, A.I., et BELLHOUSE, D.R. (1983). A convenient algorithm for drawing a simple random sampling. *Applied Statistics*, 32, 182-184.

SUNTER, A.B. (1977). List sequential sampling with equal or unequal probabilities without replacement. *Applied Statistics*, 26, 261-268.

SUNTER, A.B. (1986). Solutions to the problem of unequal probability sampling without replacement. *Revue Internationale de Statistique*, 54, 33-50.

YATES, F., et GRUNDY, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society, B*, 15, 235-261.



Tableaux, majorants du biais et simulations

ANNEXE 2

Tableau 1  
Valeur des majorants du biais  $C_\alpha$

$N$	$n$	$M = \frac{n}{N}$	$M = \frac{n}{2N}$	$M = \frac{n}{3N}$	$M = \frac{n}{4N}$	$M = \frac{n}{5N}$
Valeur du coefficient $C_\alpha$						
100	50	0.000000	0.000000	0.000000	0.000000	0.000000
	25	0.057326	0.000185	0.000015	0.000235	0.000002
	12	0.041716	0.002604	0.000235	0.000235	0.000002
	6	0.032227	0.002029	0.000134	0.000005	0.000000
	3	0.023515	0.000645	0.000000		
	250	0.000000	0.000000	0.000000	0.000000	0.000000
	125	0.129091	0.006002	0.000437	0.000038	0.000004
	62	0.090863	0.005664	0.000534	0.000059	0.000007
	31	0.066891	0.004666	0.000484	0.000059	0.000008
	15	0.048544	0.003586	0.000384	0.000046	0.000006
2,500	1,250	0.000000	0.000000	0.000000	0.000000	0.000000
	625	0.289060	0.013495	0.000987	0.000086	0.000008
	312	0.202458	0.012607	0.001190	0.000133	0.000016
	156	0.147113	0.010234	0.001064	0.000130	0.000017
	78	0.105662	0.007742	0.000841	0.000107	0.000015
	39	0.075975	0.005719	0.000634	0.000082	0.000012
	19	0.054525	0.004174	0.000466	0.000060	0.000008
	9	0.039560	0.003014	0.000301	0.000029	0.000002
	4	0.028388	0.001451	0.000034	0.000000	
	3,125	0.646539	0.030208	0.002211	0.000193	0.000018
12,500	1,562	0.452450	0.028177	0.002261	0.000297	0.000036
	781	0.327879	0.022298	0.002371	0.000290	0.000039
	390	0.234114	0.017131	0.001863	0.000238	0.000033
	195	0.166626	0.012500	0.001388	0.000181	0.000026
	97	0.118377	0.008995	0.001009	0.000133	0.000019
	48	0.084217	0.006452	0.000727	0.000096	0.000014
	24	0.060797	0.004689	0.000529	0.000069	0.000010
	12	0.044677	0.003461	0.000377	0.000044	0.000005
	6	0.033727	0.002356	0.000173	0.000008	0.000000
	3	0.024172	0.000712	0.000000		
312,500	3,906	0.732684	0.050942	0.005299	0.000649	0.000087
	1,953	0.522918	0.038250	0.004159	0.000531	0.000074
	976	0.371301	0.027833	0.003092	0.000403	0.000057
	488	0.263300	0.019979	0.002243	0.000295	0.000042
	244	0.186736	0.014259	0.001609	0.000213	0.000031
	122	0.132633	0.010168	0.001150	0.000152	0.000022
	61	0.094601	0.007273	0.000823	0.000109	0.000016
	30	0.067467	0.005027	0.000590	0.000078	0.000011
	15	0.049227	0.003820	0.000427	0.000054	0.000007
	7	0.035847	0.002637	0.000227	0.000016	0.000001
4,882	4,882	0.829762	0.062191	0.006909	0.000901	0.000128
	2,441	0.587909	0.044596	0.005006	0.000659	0.000095
	1,220	0.416165	0.031758	0.003583	0.000474	0.000068
	610	0.294647	0.022555	0.002551	0.000339	0.000049
	305	0.208743	0.016008	0.001813	0.000241	0.000035
	152	0.147877	0.011566	0.001287	0.000171	0.000025
	76	0.105272	0.008098	0.000918	0.000122	0.000018
	38	0.075422	0.005817	0.000659	0.000087	0.000013
	19	0.054695	0.004238	0.000479	0.000062	0.000009
	9	0.039644	0.003038	0.000305	0.000030	0.000002
312,500	4,882	0.829762	0.062191	0.006909	0.000901	0.000128
	2,441	0.587909	0.044596	0.005006	0.000659	0.000095
	1,220	0.416165	0.031758	0.003583	0.000474	0.000068
	610	0.294647	0.022555	0.002551	0.000339	0.000049
	305	0.208743	0.016008	0.001813	0.000241	0.000035
	152	0.147877	0.011566	0.001287	0.000171	0.000025
	76	0.105272	0.008098	0.000918	0.000122	0.000018
	38	0.075422	0.005817	0.000659	0.000087	0.000013
	19	0.054695	0.004238	0.000479	0.000062	0.000009
	9	0.039644	0.003038	0.000305	0.000030	0.000002

$\rho^2$	Plan	Paramètres	$E_{sim} \text{Var} \hat{y}$	$\text{Var} \hat{y}$	$\widehat{EQM}_{sim \hat{y}}$
0.0	<i>sm</i>	$M = 18.83N/n$	0.01318	0.01317	0.01301
	<i>srs</i>		0.01317	0.01316	0.01296
	<i>strat</i>	$H = 2$	0.01319	0.01319	0.01318
	<i>sm</i>	$M = 18.83N/n$	0.01210	0.01210	0.01187
	<i>srs</i>		0.01316	0.01316	0.01287
	<i>strat</i>	$H = 2$	0.01172	0.01188	0.01164
	<i>sm</i>	$M = 18.83N/n$	0.01073	0.01073	0.01080
	<i>srs</i>		0.01316	0.01316	0.01320
	<i>strat</i>	$H = 2$	0.00943	0.00929	0.00946
	<i>sm</i>	$M = 18.83N/n$	0.00957	0.00957	0.00954
0.6	<i>sm</i>	$M = 18.83N/n$	0.00957	0.00957	0.00954
	<i>srs</i>		0.01315	0.01316	0.01301
	<i>strat</i>	$H = 2$	0.00783	0.00778	0.00774
	<i>sm</i>	$M = 18.83N/n$	0.00839	0.00839	0.00839
	<i>srs</i>		0.01315	0.01316	0.01322
	<i>strat</i>	$H = 2$	0.00630	0.00624	0.00622
	<i>sm</i>	$M = 18.83N/n$	0.00757	0.00757	0.00760
	<i>srs</i>		0.01314	0.01316	0.01319
	<i>strat</i>	$H = 2$	0.00514	0.00508	0.00513
	<i>sm</i>	$M = 18.83N/n$	0.01319	0.01319	0.01317
0.0	<i>sm</i>	$M = 8.65N/n$	0.01317	0.01317	0.01316
	<i>srs</i>		0.01321	0.01324	0.01325
	<i>strat</i>	$H = 8$	0.01067	0.01067	0.01046
	<i>sm</i>	$M = 4.21N/n$	0.01316	0.01316	0.01287
	<i>srs</i>		0.01316	0.01316	0.01287
	<i>strat</i>	$H = 8$	0.01055	0.01047	0.01025
	<i>sm</i>	$M = 4.21N/n$	0.00810	0.00809	0.00808
	<i>srs</i>		0.01316	0.01316	0.01320
	<i>strat</i>	$H = 8$	0.00794	0.00789	0.00789
	<i>sm</i>	$M = 4.21N/n$	0.00592	0.00592	0.00588
0.6	<i>sm</i>	$M = 4.21N/n$	0.00592	0.00592	0.00588
	<i>srs</i>		0.01315	0.01316	0.01301
	<i>strat</i>	$H = 8$	0.00575	0.00564	0.00561
	<i>sm</i>	$M = 4.21N/n$	0.00344	0.00344	0.00345
	<i>srs</i>		0.01315	0.01316	0.01322
	<i>strat</i>	$H = 8$	0.00315	0.00311	0.00308
	<i>sm</i>	$M = 4.21N/n$	0.00124	0.00124	0.00125
	<i>srs</i>		0.01314	0.01316	0.01319
	<i>strat</i>	$H = 8$	0.00085	0.00079	0.00080
	<i>sm</i>	$M = 4.21N/n$	0.00012	0.00012	0.00012
1.0	<i>sm</i>	$M = 4.21N/n$	0.00012	0.00012	0.00012
	<i>srs</i>		0.01314	0.01316	0.01319
	<i>strat</i>	$H = 4$	0.00206	0.00197	0.00197
	<i>sm</i>	$M = 8.65N/n$	0.00312	0.00312	0.00313
	<i>srs</i>		0.01316	0.01316	0.01319
	<i>strat</i>	$H = 4$	0.00402	0.00391	0.00390
	<i>sm</i>	$M = 8.65N/n$	0.00484	0.00484	0.00485
	<i>srs</i>		0.01315	0.01316	0.01322
	<i>strat</i>	$H = 4$	0.00484	0.00484	0.00485
	<i>sm</i>	$M = 8.65N/n$	0.00695	0.00694	0.00688

Résultats des simulations, plan simple, stratification et stratification mobile

Tableau 2

## ANNEXE 1

## Démonstration des lemmes et des propositions

## Démonstration du lemme 3

$$\text{Var}[n_{i+1}]$$

$$= \text{Var}[n_i] + \text{Var}[I_{i+1}]$$

$$+ 2E\left(E\left\{\left(n_i - i \frac{N}{n}\right)E\left[I_{i+1} - \frac{N}{n} \mid n_i\right]\right\}\right).$$

Comme

$$2E\left[E\left\{\left(n_i - i \frac{N}{n}\right)E\left[I_{i+1} - \frac{N}{n} \mid n_i\right]\right\}\right]$$

$$= 2E\left[\left(n_i - i \frac{N}{n}\right)\left(\frac{b_i}{b_i + iN/N - n_i} - \frac{N}{n}\right)\right]$$

$$= -\frac{b_i}{2} \text{Var}[n_i],$$

on obtient

$$\text{Var}[n_{i+1}] = \text{Var}[n_i] \frac{b_i}{b_i - 2} + \frac{N}{n} \frac{N}{N - n},$$

$$(18) \quad i = 1, \dots, N - 1.$$

On montre ensuite que (3) vérifie l'équation de récurrence (18) et la condition initiale donnée par

$$\text{Var}(n_1) = \frac{N}{n} \frac{N}{N - n}.$$

## Démonstration de la proposition 1

Cas 1:  $i = 0$ . Par le lemme 2, on a directement:

$$E[I_k I_1] = E[E[I_k \mid n_1] n_1]$$

$$= \frac{N^2}{n^2} - \frac{N}{n} \frac{N}{N - n} \frac{1}{b_{k-1}} \prod_{\ell=1}^{k-2} \frac{b_\ell}{b_\ell - 1}.$$

A partir de l'étape  $N - M$ , l'algorithme est un algorithme de sélection-rejet tel qu'il est décrit dans la section 3.1. Cet algorithme aboutit au prélèvement de exactement  $n - n_{N-M}$  unités d'observation durant les  $M$  dernières étapes. Comme  $n - n_{N-M} \leq M$ , cette opération ne pose pas de problème et l'algorithme est donc de taille fixe  $n$ .

$$\text{Pr}[0 \leq n - n_{N-M} \leq M] = 1.$$

Donc,

$$\text{Pr}\left[n - M - \frac{N}{n} < n_{N-M} < \frac{N}{N - n} + n\right] = 1.$$

Par (13), on a

## Démonstration de la proposition 2

Or, le lemme 3 nous donne  $\text{Var}[n_i]$ . On obtient direc-

tement (4).

$$= \frac{N^2}{n^2} - \frac{1}{b_{i+k-1}} \left\{ \frac{N}{n} \frac{N}{N - n} - \frac{b_i}{\text{Var}[n_i]} \right\} \prod_{\ell=i+1}^{i+k-2} \frac{b_\ell}{b_\ell - 1}.$$

$$\times \left\{ \frac{N}{n} - \left( n_i - i \frac{N}{n} \right) \frac{1}{b_i} \right\}$$

$$= E\left\{ \frac{N}{n} - \left( n_i + 1 - (i + 1) \frac{N}{n} \right) \frac{1}{b_{i+k-1}} \prod_{\ell=i+1}^{i+k-2} \frac{b_\ell}{b_\ell - 1} \right\}$$

$$E[E[I_{i+k} I_{i+1} \mid n_i]]$$

Ce qui donne

$$\times \left\{ \frac{N}{n} - \left( i - i \frac{N}{n} \right) \frac{1}{b_i} \right\}.$$

$$= \left\{ \frac{N}{n} - (i + 1) \frac{N}{n} \right\} \frac{1}{b_{i+k-1}} \prod_{\ell=i+1}^{i+k-2} \frac{b_\ell}{b_\ell - 1}$$

$$= E[I_{i+k} \mid n_{i+1} = i + 1] E[I_{i+1} \mid n_i = i]$$

$$E[I_{i+k} I_{i+1} \mid n_i = i]$$

Cas 2:  $i > 0$ . En utilisant le lemme 2, on obtient:



Une lecture attentive des résultats semble indiquer que l'estimateur de variance proposé pour l'algorithme de la strate mobile ne souffre pas d'un biais systématique quelle que soit la valeur du coefficient de corrélation entre  $x$  et  $y$ . Les résultats semblent également indiquer que l'expression approximative donnée pour la variance de l'estimateur de moyenne pour la stratification mobile est une approximation valable.

4.5 Intérêt de l'algorithme

Dans la classe des algorithmes définis par l'algorithme général, on appelle horizon moyen d'un algorithme, la quantité

$$b = \frac{1}{N-1} \sum_{i=0}^I b_i.$$

Pour le plan simple, on obtient  $b_{sr} = (N + 1)/2$ . Pour l'algorithme de la strate mobile, on a

$$\bar{b}_{sm} = \frac{1}{N-M-1} \sum_{i=N-M-1}^I M + \sum_{i=N-1}^{i=N-M} (N-i)$$

$$= \frac{M}{N} \left\{ N - \frac{M-1}{2} \right\}.$$

Supposons maintenant que nous sélectionnions, comme décrit en section 3.2, un échantillon au moyen d'un plan avec allocation proportionnelle où toutes les strates ont la même taille et où les tailles des  $H$  strates sont toutes égales. Dans un tel plan, l'horizon moyen vaut

$$b_{srat} = \frac{1}{N} \left( \frac{H}{2} + 1 \right).$$

Un changement d'horizon moyen n'affecte pas fondamentalement les probabilités d'inclusion d'ordre un. Les probabilités d'inclusion d'ordre deux sont par contre fortement modifiées par un changement d'horizon. En effet, on voit aisément que plus l'horizon moyen est petit, plus la probabilité de sélectionner deux individus proches est faible. (On dit que deux individus sont proches si la valeur absolue de la différence de leur numéro d'ordre sur le fichier de données est petite.) Intuitivement, on peut donc s'attendre à ce que l'algorithme de la strate mobile ait un effet de stratification similaire à un plan stratifié avec allocation proportionnelle ayant le même horizon moyen, c'est-à-dire quand

$$b_{srat} = b_{sm},$$

L'auteur remercie Pierre Lavallée pour les conseils qu'il a exprimés sur des versions antérieures de cet article. L'auteur tient également à remercier un éditeur associé et un arbitre qui ont émis de nombreux commentaires constructifs qui ont permis d'améliorer considérablement cet article.

REMERCIEMENTS

Les simulations effectuées montrent clairement que l'algorithme de stratification mobile donne un effet de stratification du même type qu'une stratification classique avec allocation proportionnelle. Cet algorithme permet d'étudier le délicat problème du découpage d'une variable continue en strates. Les estimateurs de moyenne proposés sont légèrement biaisés. Cependant, dès que  $M \geq 10N/n$ , les simulations montrent qu'il est extrêmement rare qu'au moins un des  $c_i$  soit en dehors de  $[0, 1]$ . De plus, nous avons montré que même quand cette probabilité n'est pas nulle, le biais de l'estimateur que nous proposons est négligeable dès que  $M \geq 3N/n$ .

5. REMARQUES

Pour chaque ensemble de simulations présenté en annexe (tableau 2), les tailles des strates mobiles (cas:  $sm$ ) ont été fixées en fonction du nombre de strates (cas:  $srat$ ) de manière à ce que les horizons moyens des deux plans soient identiques au sens de l'expression (17). On constate que, dans ce cas, le gain de précision (comparativement au plan simple) obtenu au moyen de l'algorithme de la strate mobile est du même ordre de grandeur qu'au moyen de la stratification.

$$M \approx \frac{H}{2N}.$$

Quand  $N$  est grand par rapport à  $M$ , on a approximativement

$$M = N + \frac{1}{2} - \sqrt{\frac{1}{4} + N^2 \frac{H-1}{H}}. \tag{17}$$

ou autrement dit, quand

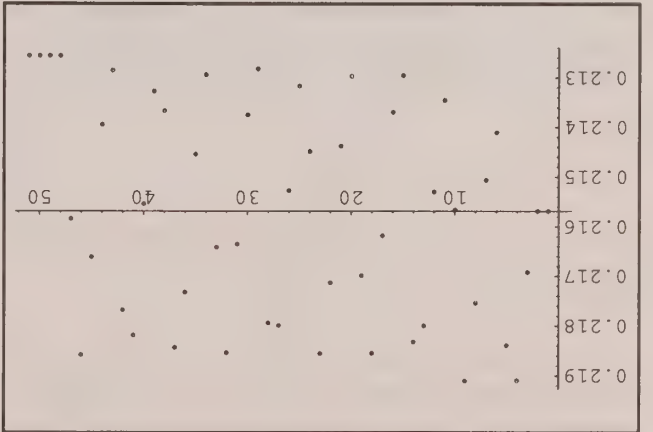


Figure 1. Probabilités d'inclusion.

de  $n/N$  sans qu'il y ait une tendance notable liée à l'ordre du fichier. En pratique, on peut considérer qu'il est très peu probable d'avoir une variable d'inclusion fortement corrélée aux probabilités d'inclusion exactes et le biais sera donc, le plus souvent, nettement inférieur au majorant donné.

On pourrait, bien sûr, utiliser les probabilités d'inclusion exactes pour procéder à une estimation. Nous pensons que cela n'en vaut pas la peine pour deux raisons:

- d'abord parce que le calcul des probabilités d'inclusion exactes nécessite un temps de calcul important,
- ensuite parce que les probabilités d'inclusion d'ordre un exactes sont telles que

$$\text{Var} \left[ \sum_{i \in S} \frac{1}{\pi_i} \right] \neq 0.$$

Dans ce cas, l'estimateur de Horvitz-Thompson d'une variable constante ( $y_k = C$ ) est aléatoire. Pour pallier ce problème, on estime généralement la moyenne par le ratio de Hájek (1971). Cet estimateur est également biaisé.

#### 4.4 Estimation de la variance de l'estimateur

En supposant que  $\Pr(0 \leq c_i \leq 1) \approx 1$ , on peut construire également une approximation des probabilités d'inclusion d'ordre deux à partir de corollaire 1. En considérant que  $b_i$  vaut  $M$  si  $i \leq N - M$  et  $N - i$  sinon, on obtient l'approximation suivante:

$$\pi_{ik} \approx \frac{n^2}{N^2} (1 - \theta_{ik})$$

ou

de l'approximation proposée.

- les variances des estimateurs de moyenne. Ces variances sont données par les expressions (9) (srs) et (15) (sm). Dans le cas de la stratification mobile, il s'agit bien sûr de l'approximation proposée.
  - les erreurs quadratiques moyennes sur les simulations des estimateurs de moyenne. Ces quantités sont notées  $EQM_{sm}(\hat{y}) = E_{sm}(\hat{y} - y)^2$ .
  - les moyennes sur les simulations des estimateurs de variance de l'estimateur de la moyenne qui sont notées  $E_{sm} \widehat{\text{Var}}(\hat{y})$ . Ces estimateurs de variance sont donnés par les expressions (11) (srs) et (16) (sm).
- Pour chacune des simulations, on donne 3 résultats:

200,000 échantillons ont été sélectionnés.

cation sur les choix de  $M$  et  $H$ . Pour chaque simulation, strates  $H$  (cas: *strat*). Nous donnerons plus loin une expli- la strate mobile  $M$  (cas: *sm*) et pour différents nombres de simulations ont été effectuées pour différentes valeurs de oeuvre en utilisant les mêmes nombres aléatoires. Des cas particuliers de l'algorithme général et ont été mises en plan simple sans remise (srs). Ces trois méthodes sont des où les tailles des strates sont toutes égales (*strat*) et d'un (sm), d'un plan stratifié avec allocation proportionnelle été sélectionnés au moyen de la méthode de la strate mobile Dans ces populations, des échantillons de taille 64 ont en fonction de la variable  $x$ . L'objectif est d'estimer  $y$ .

de corrélation  $\rho$  fixé. Ensuite, les populations ont été triées ayant une distribution normale bivariable avec un coefficient ont été générées au moyen de nombres pseudo-aléatoires  $N = 400$ . Les valeurs prises par les deux variables  $x$  et  $y$  en annexe. Nous avons généré des populations de taille simulations. Les résultats sont donnés dans le tableau 2 l'ampleur de ce biais, nous avons effectué un ensemble de De nouveau, cet estimateur est biaisé. Afin de juger de

$$\widehat{\text{Var}}_{app}[\hat{y}_{sm}] = \frac{1}{N^2} \sum_{i \in S} \sum_{k \neq i} (y_i - y_k)^2 \frac{1}{\theta_{ik}}. \quad (16)$$

variance de l'estimateur de moyenne:

$$\text{Var}_{app}[\hat{y}_{sm}] = \frac{1}{N^2} \sum_{i \in U} \sum_{k \in U, k \neq i} (y_i - y_k)^2 \theta_{ik}. \quad (15)$$

variance de  $\hat{y}_{sm}$ :

En supposant que les probabilités d'inclusion d'ordre un valent  $n/N$ , on peut construire une approximation de la

$$\theta_{ik} = \frac{N - n}{N - n} \frac{2n}{M - 1} \left\{ 1 + \left( \frac{M}{M - 2} \right)^{\min(i-1, N-M)} \right\} \times \left( \frac{M - 1}{\max(0, \min(N - M - i + 1, k - i))} \right) \quad k > i.$$



valeur supérieure à 1 est cependant toujours inférieure à  $1 + n/(NM)$ . En effet, dans le cas où un des  $c_i$  est déjà supérieur à 1, l'unité  $i$  est toujours sélectionnée et donc  $c_{i+1}$  prend une valeur inférieure à  $c_i$ .

On obtient

$$\Pr \left[ -\frac{N-n}{NM} < c_i < 1 + \frac{NM}{n} \right] = 1, i = 0, \dots, N-M. \quad (13)$$

Le plan est cependant de taille fixe, ce résultat est donné par la proposition suivante:

**Proposition 2** Si  $b_i = \min(M, N-i)$ ,  $(N/n < M < N)$ ,  $0 = 1, \dots, N-1$ , alors le plan est de taille fixe.

La démonstration est donnée en annexe.

Comme les  $c_i$  ne sont pas toujours dans l'intervalle

$[0, 1]$ , nous avons effectué 50 simulations d'application de l'algorithme de la strate mobile pour différentes tailles de population et d'échantillon. Les tailles de population  $N$  choisies sont 100, 500, 2500, 12500, 62500, 312500. Les inverses des taux de sondages  $(N/n)$  sont 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096. Nous avons effectué plusieurs simulations en faisant varier la taille de la strate mobile ainsi:  $M = N/n, 2N/n, 3N/n, \dots$ . Les simulations semblent indiquer que plus  $M$  est grand, plus la probabilité qu'un  $c_i$  soit en dehors de  $[0, 1]$  est faible.

Dès que  $M \geq 10N/n$ , pour toutes les simulations que nous avons effectuées, le problème ne s'est plus jamais posé. Ce premier résultat n'implique pas que la probabilité qu'au moins un des  $c_i$  soit hors de  $[0, 1]$  est nulle quand  $M \geq 10N/n$ . Cependant on peut penser que, dans ce cas, cette probabilité est très faible.

#### 4.3 Estimation de la moyenne et biais

En considérant les résultats donnés par l'expression (2) et la proposition 1, on obtient, en première approximation, que les probabilités d'inclusion d'ordre un valent approximativement  $\pi_i \approx n/N$ . Cette approximation des probabilités d'inclusion permet de construire un estimateur.

$$\hat{y}_{sm} = \frac{1}{n} \sum_{i \in U} y_i.$$

Cet estimateur est légèrement biaisé car les  $c_i$  ne sont pas tous exactement dans l'intervalle  $[0, 1]$ . Ce biais vaut

$$B[\hat{y}_{sm}] = \frac{1}{n} \sum_{i \in U} \alpha_i y_i$$

où  $\alpha_i = \pi_i N/n - 1$ . Comme le plan est de taille fixe,  $\sum_{i \in U} \alpha_i = 0$ . On peut donc écrire le biais sous la forme d'une covariances:  $B[\hat{y}_{sm}] = \sigma_{y\alpha}$  où

$$\sigma_{y\alpha} = \frac{1}{n} \sum_{i \in U} \alpha_i (y_i - \bar{y}). \quad (14)$$

Comme la valeur absolue d'une covariance est toujours inférieure ou égale au produit des deux écart-types, on obtient un majorant pour la valeur absolue du biais

$$|B[\hat{y}_{sm}]| \leq \sigma_y \sigma_\alpha$$

où  $\sigma_y$  est défini par (10) et

$$\sigma_\alpha^2 = \frac{1}{N} \sum_{i \in U} \alpha_i^2.$$

La variance de l'estimateur est d'un ordre de grandeur comparable (pour  $N$  et  $n$  fixé) à la variance de l'estimateur de la moyenne dans le plan simple sans remise. On peut donc écrire

$$|B[\hat{y}_{sm}]| \leq C_\alpha \sqrt{\text{Var}[\hat{y}_{srs}]} \quad \text{ou } \text{Var}[\hat{y}_{srs}] \text{ est défini par (9) et}$$

$$C_\alpha = \sigma_\alpha \sqrt{\frac{n(N-1)}{N-n}}.$$

Nous allons considérer que le biais est négligeable quand le majorant du biais de l'estimateur  $\hat{y}_{sm}$  est négligeable par rapport à  $\text{Var}[\hat{y}_{srs}]^{1/2}$ , c'est-à-dire quand  $C_\alpha$  est petit.

On peut calculer récursivement la valeur exacte des

$$\Pr[I_i = 1 | n_i] = \bar{c}_i, i = 1, \dots, N-M$$

où  $\bar{c}_i$  vaut 0 si  $c_i < 0$ ,  $c_i$  si  $0 \leq c_i \leq 1$  et 1 si  $c_i > 1$ . De ce résultat, on peut déduire la valeur exacte des probabilités d'inclusion d'ordre un.

Nous avons calculé (annexe, tableau 1) les valeurs de  $C_\alpha$  pour diverses tailles de population (100 à 312500) et d'échantillon. Les valeurs de  $C_\alpha$  sont données pour des tailles de strates mobiles  $M$  égales à  $N/n, 2N/n, 3N/n, 4N/n$  et  $5N/n$ . On voit que dès que la strate mobile vaut  $2N/n$ ,  $C_\alpha$  ne dépasse jamais 0.07. Quand  $M = 3N/n$ , le coefficient  $C_\alpha$  s'exprime en millièmes. Selon Cochran (1977, pp. 13-14), le biais peut alors être négligé. Ce tableau montre donc que si  $M \geq 3N/n$ , le biais de l'estimateur sera négligeable au moins pour les tailles des populations et des échantillons spécifiés.

Ces résultats n'impliquent cependant pas que le biais de l'estimateur est grand quand  $M$  est très petit (par exemple  $M = N/n$ ). Les  $C_\alpha$  sont des majorants des biais. Par l'expression (14), on voit que le biais sera d'autant plus grand que la variable d'intérêt est corrélée aux probabilités d'inclusion exactes. Nous avons représenté (figure 1), les probabilités d'inclusion exactes (en ordonnée) pour les  $N$  individus (en abscisse) obtenues par l'algorithme de stratification mobile avec les paramètres  $N = 51, n = 11, M = N/n$ . Ce cas est évidemment très défavorable. Le résultat est intéressant. Dans ce cas,  $n/N = 0.215686$ . Les probabilités d'inclusion se répartissent de part et d'autre

$$(8) \quad \hat{y}_{srs} = \frac{1}{n} \sum_{i \in s} y_i.$$

La variance de cet estimateur est donnée par

$$(9) \quad \text{Var}[\hat{y}_{srs}] = \frac{\sigma_y^2}{n} \frac{N-1}{N}.$$

où

$$(10) \quad \sigma_y^2 = \frac{1}{N} \sum_{i \in U} (y_i - \bar{y})^2.$$

Cette variance peut être estimée sans biais par

$$(11) \quad \widehat{\text{Var}}[\hat{y}_{srs}] = \frac{s_y^2}{n} \frac{N-1}{N}$$

où

$$(12) \quad s_y^2 = \frac{1}{n-1} \sum_{i \in s} (y_i - \hat{y}_{srs})^2.$$

### 3.2 Plan stratifié

On peut définir également le plan stratifié au moyen de l'algorithme général. La variable de stratification est ici le numéro d'ordre de l'individu. Considérons le cas partiel d'un plan stratifié en  $H$  strates avec allocation proportionnelle où toutes les strates ont la même taille. Les strates sont constituées de telle manière que les individus d'une même strate soient contigus sur le fichier de données. On suppose, en outre, que  $N/H$  est entier. Ce plan stratifié s'obtient en prenant simplement

$$b_i = \left\{ (N - i - 1) \bmod \frac{H}{N} + 1, i = 0, \dots, N-1. \right.$$

## 4. APPLICATION 2: STRATIFICATION MOBILE

### 4.1 Le problème

Le fichier est supposé être ordonné selon une variable auxiliaire proche de la variable d'intérêt. Le problème est le suivant: Comment effectuer un tirage aléatoire qui donne une petite variance pour l'estimateur de Horvitz-Thompson d'une moyenne? En examinant la formulation de la variance de Yates-Grundy (5), on voit que l'on peut répondre à cette question de deux manières bien distinctes. La première solution consiste à tirer à probabilités inégales avec des probabilités d'inclusion d'ordre un proportionnelles à la variable d'intérêt. S'il est possible d'effectuer un tel tirage, toutes les quantités

$$\left( y_i - \frac{\pi_i}{\pi_k} \right)^2$$

seraient nulles et donc la variance serait nulle.

La seconde solution consiste à jouer sur les probabilités d'inclusion d'ordre deux. Un bon tirage pourrait être tel que les  $\pi_{ik}$  soient proches de  $\pi_i \pi_k$  si  $y_i$  est très différent de  $y_k$ . D'autre part, si  $y_i$  est très proche de  $y_k$ , on peut choisir des probabilités d'inclusion d'ordre deux  $\pi_{ik}$  nettement inférieures à  $\pi_i \pi_k$ . Ainsi, quand les quantités

$$\left( y_i - \frac{\pi_i}{\pi_k} \right)^2$$

seraient grandes (resp. petites), les quantités  $\pi_i \pi_k - \pi_{ik}$  seraient petites (resp. grandes). On obtiendrait donc une petite variance.

La seconde solution que nous venons de décrire est en fait déjà abondamment utilisée. C'est l'idée de base de la stratification. Notre objectif est d'appliquer cette idée pour construire un algorithme séquentiel de tirage facile à mettre en oeuvre. Cet algorithme pourrait s'appliquer à n'importe quel fichier sans qu'il soit nécessaire de connaître autre chose que la taille de la population. Il serait donc applicable à de très gros fichiers. On profiterait ainsi de l'information apportée par cette variable auxiliaire comme pour une stratification, sans pour autant devoir se préoccuper d'un réel découpage en strates.

### 4.2 La méthode

On définit d'abord  $M$ , la longueur de la strate mobile dans la population.  $M$  représente en quelque sorte la taille de la strate dans la population et est telle que  $N/n \leq M \leq N$ . L'algorithme de la strate mobile est défini par

$$b_i = \min(M, N - i), i = 0, \dots, N-1.$$

Un problème se pose cependant. Les quantités  $c_i$  définies par

$$c_i = \begin{cases} \frac{(M+i)n/N - n_i}{M} & \text{si } i \leq N-M \\ \frac{n - n_i}{N-i} & \text{sinon,} \end{cases}$$

ne sont pas toujours dans  $[0, 1]$ .

En effet, supposons que, avant la  $(N-M)$ -ième étape de l'algorithme,  $c_i$  soit positif et très proche de 0 et que par malchance l'unité  $i$  soit malgré cela choisie. Dans ce cas,  $c_{i+1}$  prendrait la valeur  $c_i - (N-n)/(NM)$ .  $c_{i+1}$  peut donc prendre une valeur négative mais cette valeur négative est toujours supérieure à  $-(N-n)/(NM)$ . En effet, dans le cas où un des  $c_i$  est déjà négatif, l'unité  $i$  n'est pas sélectionnée et donc  $c_{i+1}$  prend une valeur supérieure à  $c_i$ . Supposons maintenant que avant la  $(N-M)$ -ième étape de l'algorithme, un  $c_i$  soit très légèrement inférieur à 1 et que malgré cela l'unité  $i$  ne soit pas sélectionnée. Dans ce cas,  $c_{i+1}$  prendrait la valeur  $c_i + n/(NM)$ .  $c_{i+1}$  peut donc prendre une valeur plus grande que 1 mais cette



**Lemme 2** Si  $\Pr[0 \leq c_i \leq 1] = 1, i = 0, \dots, N-1$ , alors

$$E[I_{i+k} | n_i]$$

$$= \frac{N}{n} - i \left( \frac{N}{n} \right) \frac{1}{1} \prod_{\ell=k-2}^{i+k-1} \frac{b_{i+\ell}}{b_{i+\ell}-1},$$

$$i = 1, \dots, N-1, k = 1, \dots, N-i.$$

On suppose par convention qu'un produit vide vaut 1. **Lemme 3** Si  $\Pr[0 \leq c_i \leq 1] = 1, i = 0, \dots, N-1$ , alors

$$\text{Var}[n_i] = \frac{N}{n} \frac{N}{N-n} \sum_{i=1}^{j=1} \prod_{\ell=1}^{i+k-2} \frac{b_{i+\ell}}{b_{i+\ell}-2}, i = 1, \dots, N. \quad (3)$$

La démonstration est donnée en annexe.

Enfin, la probabilité d'inclusion d'ordre deux est donnée par la proposition suivante:

**Proposition 1** Si  $\Pr[0 \leq c_i \leq 1] = 1, i = 0, \dots, N-1$ , alors

$$E[I_{i+k} I_{i+1}]$$

$$= \frac{N^2}{n^2} - \frac{N}{n} \frac{N}{N-n} \frac{1}{b_{i+k-1}}$$

$$\times \left( 1 - \frac{1}{b_i} \sum_{j=1}^{i+k-2} \prod_{\ell=1}^{j-1} \frac{b_{i+\ell}}{b_{i+\ell}-2} \right) \prod_{\ell=i+1}^{i+k-2} \frac{b_{i+\ell}}{b_{i+\ell}-1},$$

$$i = 0, \dots, N-2, k = 2, \dots, N-i. \quad (4)$$

La démonstration est donnée en annexe.

**Corollaire 1** Si  $\Pr[0 \leq c_i \leq 1] = 1, i = 0, \dots, N-1$ , alors

$$\pi_{ik} = \frac{N^2}{n^2} - \frac{N}{n} \frac{N}{N-n} \left( 1 - \frac{1}{b_{i-1}} \sum_{j=1}^{i-2} \prod_{\ell=1}^{j-1} \frac{b_{i-\ell}}{b_{i-\ell}-2} \right) \prod_{\ell=i-2}^{i-1} \frac{b_{i-\ell}}{b_{i-\ell}-1}$$

$$\times \frac{1}{b_{i-2}} \prod_{\ell=i-2}^{i-1} \frac{b_{i-\ell}}{b_{i-\ell}-1}, i = 1, \dots, N-1, k > i.$$

## 2.4 L'estimateur de Horvitz-Thompson et sa variance

L'estimateur de Horvitz-Thompson est la moyenne d'échantillon simple puisque les probabilités d'inclusion d'ordre un sont toutes égales

$$\hat{y}_{\pi} = \frac{1}{n} \sum_{i \in S} y_i.$$

Rappelons également quelques résultats classiques concernant le plan simple que nous utiliserons plus loin. L'estimateur de  $y$  est donc la moyenne de l'échantillon

$$\pi_{ik} = \frac{n(n-1)}{N(N-1)}.$$

On a toujours  $0 \leq c_i \leq 1$ . Les probabilités d'inclusion d'ordre un valent toujours  $n/N$ . Le calcul des probabilités d'inclusion d'ordre deux découle de la proposition 1. En supposant  $k > i$ , par le corollaire 1, on peut retrouver les probabilités d'inclusion d'ordre deux du plan simple:

$$b_i = N - i, i = 0, \dots, N-1.$$

L'algorithme de tirage le plus simple, la méthode de sélection-rejet telle qu'elle est décrite dans Fan, Fuller et Deville et Grosbras (1987, p. 210) est bien sûr un cas particulier de l'algorithme général. Il suffit de prendre

## 3.1 Plan simple

## 3. APPLICATION 1: TIRAGES ALÉATOIRES SIMPLE ET STRATIFIÉ

$$\widehat{\text{Var}}[\hat{y}_{\pi}] = \frac{1}{2n^2} \sum_{i \in S} \sum_{\substack{k \in S \\ k \neq i}} (y_i - y_k)^2 \frac{1 - \gamma_{ik}}{\gamma_{ik}}.$$

Ce qui peut s'écrire ici

$$\widehat{\text{Var}}[\hat{y}_{\pi}] = \frac{1}{2N^2} \sum_{i \in S} \sum_{\substack{k \in S \\ k \neq i}} \left( \frac{\pi_i}{y_i} - \frac{\pi_k}{y_k} \right)^2 \frac{\pi_{ik}}{\pi_i \pi_k - \pi_{ik}}. \quad (7)$$

L'estimateur de la variance est donné par

$$\text{Var}[\hat{y}_{\pi}] = \frac{1}{N^2} \sum_{i \in U} \sum_{\substack{k \in U \\ k \neq i}} (y_i - y_k)^2 \gamma_{ik}. \quad (6)$$

on peut écrire

$$\gamma_{ik} = 1 - \frac{n^2}{N^2},$$

Comme  $\pi_i = n/N, i = 1, \dots, N$ , et en posant

$$\text{Var}[\hat{y}_{\pi}] = \frac{1}{2N^2} \sum_{i \in U} \sum_{\substack{k \in U \\ k \neq i}} \left( \frac{\pi_i}{y_i} - \frac{\pi_k}{y_k} \right)^2 (\pi_i \pi_k - \pi_{ik}). \quad (5)$$

Si le plan est de taille fixe, on peut utiliser la formule de la variance de Yates et Grundy (1953)

soit de taille fixe ou pour que les unités soient sélectionnées à probabilités égales. Le choix de différentes valeurs pour les  $b_i$ ,  $i = 0, \dots, N-1$ , permettra de déterminer plusieurs cas particuliers de l'algorithme général.

Si les  $b_i$  sont des réels strictement positifs tel que  $b_i \leq N-i$ , alors la taille de l'échantillon est inférieure ou égale à  $n$ . En effet, supposons que l'on ait déjà prélevé  $n$  unités dans la population à l'étape  $i$  et que  $b_i \leq N-i$ , alors

$$(b_i + i)n/N - n = \frac{N}{n} - \frac{b_i}{n} \leq \frac{N}{n} - \frac{N-i}{n} = 0.$$

Il est dès lors impossible de prélever une unité supplémentaire. On supposera pour tout ce qui suit que  $b_i \leq N-i$ . Par ailleurs, si  $b_i \leq N-i$ ,  $i = 1, \dots, N-n-1$ , et que  $b_i = N-i$ ,  $i = N-n, \dots, N-1$ , alors la taille de l'échantillon est de taille fixe  $n$ . Soulignons que ces conditions pour obtenir un échantillon de taille fixe sont suffisantes mais non nécessaires.

Trois cas particuliers de l'algorithme seront étudiés plus loin. Ces trois cas sont définis par trois choix de coefficients  $b_i$ ,  $i = 0, \dots, N-1$ . Avant d'étudier ces choix particuliers, nous allons déterminer les probabilités d'inclusion d'ordre un et deux en toute généralité.

## 2.2 Probabilités d'inclusion d'ordre un

Notons  $n_i$  le nombre d'unités sélectionnées après avoir passé  $i$  enregistrements. On voit directement que  $n_1, \dots, n_i, \dots, n_N$  est une chaîne de Markov. En effet, il découle directement de l'algorithme que

$$\Pr[n_i = j \mid n_1, \dots, n_{i-1}] = \Pr[n_i = j \mid n_{i-1}].$$

Les variables aléatoires

$$c_i = \frac{b_i}{(b_i + i)n/N - n_i}, \quad i = 0, \dots, N-1,$$

peuvent parfois prendre des valeurs supérieures à 1 ou inférieures à 0. Comme  $\max(0, n - N + i) \leq n_i \leq \min(i, n)$ , on a que  $\Pr[0 \leq c_i \leq 1] = 1$  si

$$b_i \geq \begin{cases} \min\left(i \frac{N-n}{n}, N-i\right) & \text{si } n \leq N/2 \\ \min\left(i \frac{N-n}{n}, N-i\right) & \text{si } n > N/2 \end{cases},$$

$$i = 0, \dots, N-1. \quad (1)$$

Les conditions (1) sont à nouveau suffisantes mais non nécessaires. On peut donc construire des  $b_i$  qui ne satisfont pas à ces conditions mais qui fournissent des  $c_i$  dans  $[0, 1]$ . Le cas traité en section 3.2 (stratification) en donne un exemple.

## 2.3 Probabilités d'inclusion d'ordre deux

Quatre résultats donnés par les lemmes 1, 2 et 3 sont nécessaires pour déterminer les probabilités d'inclusion d'ordre deux.

**Lemme 1** Si  $\Pr[0 \leq c_i \leq 1] = 1$ ,  $i = 0, \dots, N-1$ , alors

$$E[n_{i+k} \mid n_i]$$

$$= (i+k) \frac{N}{n} + \left( n_i - i \frac{N}{n} \right) \prod_{\ell=i}^{i+k-1} \frac{b_\ell}{b_\ell - 1},$$

$$i = 1, \dots, N-1, k = 1, \dots, N-i.$$

Ce lemme se montre par récurrence en supposant qu'il est vrai pour  $k-1$ . Par le lemme 1, on obtient aisément par différence le lemme suivant:

$$E[I_{i+1} \mid n_1, \dots, n_i] = E[I_{i+1} \mid n_i] =$$

on a

$$\Pr[0 \leq c_i \leq 1] = 1, \quad i = 0, \dots, N-1,$$

Nous reviendrons sur le problème des  $c_i$  ayant des valeurs supérieures à 1 ou inférieures à 0 plus loin. Si

$$\Pr[0 \leq c_i \leq 1] = 1, \quad i = 0, \dots, N-1.$$

Pour simplifier les démonstrations qui suivent, on supposera par la suite que

condition (1).

L'exemple suivant fournit également des  $c_i$  dans  $[0, 1]$  sans respecter la condition (1): prenons  $N = 12$ ,  $n = 4$  et  $b_0 = b_1 = b_2 = b_3 = b_4 = b_5 = 7$ ,  $b_i = N-i$ ,  $i = 6, \dots, 11$ . On obtient  $c_0 = 1/3$ ,  $c_1 = (7-3n_1)/18$ ,  $c_2 = (3-n_2)/7$ ,  $c_3 = (3-n_3)/6$ ,  $c_4 = (10-3n_4)/18$ ,  $c_5 = (4-n_5)/7$ ,  $c_6 = (4-n_6)/6$ ,  $c_7 = (4-n_7)/5$ ,  $c_8 = (4-n_8)/4$ ,  $c_9 = (4-n_9)/3$ ,  $c_{10} = (4-n_{10})/2$ ,  $c_{11} = (4-n_{11})$ . Remarquons que  $n_1 \leq 1$ ,  $n_2 \leq 2$ ,  $n_3 \leq 3$ . Si  $n_3 = 3$  alors  $c_3 = 0$  et donc  $n_4 \leq 3$ . Ensuite, on obtient  $n_5 \leq 4$  et si  $n_5 = 4$  alors  $c_5 = 0$  et donc  $n_6 \leq 4$ . Cette dernière remarque est vraie pour toutes les  $c_i$  qui suivent. On constate donc que tous les  $c_i$  sont dans  $[0, 1]$  alors que  $b_4 = 6$  ne respecte pas la condition (1).



# Un algorithme de stratification mobile

YVES TILÉ<sup>1</sup>

## RÉSUMÉ

Un algorithme général à probabilités égales est présenté. On donne les probabilités d'inclusion d'ordre deux qui correspondent à cet algorithme. Celui-ci généralise la méthode de sélection-rejet permettant de constituer un échantillon au moyen d'un plan simple sans remise. Un autre cas particulier de cet algorithme baptisé «algorithme de stratification mobile» est discuté. Il permet d'obtenir un effet de stratification hissé en utilisant comme variable de stratification le numéro d'ordre des unités d'observation. On donne des approximations des probabilités d'inclusion d'ordre un et deux. Ces approximations permettent de construire un estimateur de la moyenne de la population ainsi qu'un estimateur de la variance de cet estimateur de moyenne. Cet algorithme est ensuite comparé à un plan stratifié classique avec allocation proportionnelle.

MOTS CLÉS: Algorithme de tirage; sondage à probabilités égales; strate.

## 1. INTRODUCTION

Quand un fichier est ordonné selon une variable auxi-

liaire proche de la variable d'intérêt, comment peut-on sélectionner un échantillon en tirant profit de cette information? Une solution à ce problème consiste à effectuer un tirage stratifié. Cependant, pour effectuer un tel tirage, il faut résoudre le délicat problème du découpage de la population en strates. Une autre solution simple, rapide et efficace consiste à recourir à un tirage systématique. L'algorithme s'écrit en quelques lignes. De plus, on sait que l'on profitera de la manière dont le fichier est ordonné. Le tirage systématique a cependant un défaut important: pour estimer les variances des estimateurs de totaux ou de moyennes, il est nécessaire d'énoncer une ou plusieurs hypothèses sur la population. Nous montrons qu'il existe un autre algorithme de tirage simple permettant de constituer un échantillon en un seul passage en tirant profit de l'ordonnancement du fichier. Pour cet algorithme, nous donnons un estimateur de la variance de l'estimateur d'un total ou d'une moyenne qui ne nécessite pas une modélisation de la population.

Dans la section 2, un algorithme général de tirage donnant des probabilités d'inclusion d'ordre un égales est présenté. On donne les probabilités d'inclusion d'ordre un et deux. Dans la section 3, on montre que l'algorithme présenté généralise la méthode de sélection-rejet permettant de constituer un échantillon simple sans remise et le plan stratifié avec allocation proportionnelle. Enfin, dans la section 4, on définit la méthode de la strate mobile et les conclusions sont données en section 5.

## 2. PRÉSENTATION DE L'ALGORITHME GÉNÉRAL

### 2.1 L'algorithme

Soit  $U = \{1, \dots, i, \dots, N\}$  une population finie, on note  $y_1, \dots, y_i, \dots, y_N$ , les  $N$  valeurs prises par la variable

$y$  sur les  $N$  unités d'observation de  $U$ . La moyenne des valeurs prises par la variable  $y$  sur la population est notée

$$\bar{y} = \frac{1}{N} \sum_{i \in U} y_i.$$

Un échantillon aléatoire  $s$  de taille fixe  $n$  est prélevé dans cette population. Les variables aléatoires indicatrices de la présence des unités d'observation dans  $s$  sont notées  $I_i$ ,  $i \in U$ . La probabilité d'inclusion d'ordre un est notée  $\pi_i = \Pr(i \in s) = E(I_i)$ ,  $i \in U$ , et la probabilité d'inclusion d'ordre deux  $\pi_{ik} = E(I_i I_k)$ ,  $i \neq k \in U$ . L'algorithme est très court. Il s'apparente aux algorithmes de Fan, Fuller et Rezuca (1962), Bebbington (1975), McLeod et Bellhouse (1983) et de Sunter (1977 et 1986). Il faut connaître seulement  $N$ ,  $n$  et les  $b_i$ ,  $i = 0, \dots, N - 1$ . Les autres variables sont des variables de travail.

### Algorithme général

$j > 0$ ;  
 $i > 0$ ;  
Répéter pour  $i = 0, \dots, N - 1$   
 $n > =$  un nombre aléatoire selon une loi uniforme  $[0, 1]$ ;  
si  $\frac{(b_i + i)n/N - j}{b_i} > n$  alors sélectionner l'entree  $i + 1$ ;  
sinon, passer l'enregistrement  $i + 1$ ;  
 $i > = i + 1$ .

À chaque étape,  $j$  représente le nombre d'enregistrements déjà sélectionnés et  $i$  le nombre d'enregistrements passés (sélectionnés ou non). À chaque itération, on décide de la sélection de l'enregistrement  $i + 1$ . Si cet enregistrement est sélectionné, il devient le  $(j + 1)$ -ième de l'échantillon. Les coefficients  $b_i$ ,  $i = 0, \dots, N - 1$ , sont des réels strictement positifs. Ces quantités doivent satisfaire à certaines conditions discutées plus bas pour que le plan

<sup>1</sup> Yves Tilé, Laboratoire de Méthodologie du Traitement des Données, C.P. 124, Université Libre de Bruxelles, Avenue Jeanne, 44, 1050 Bruxelles, Belgique, E-mail ytileb@ulb.ac.be





5. CONCLUSION

L'article vise à présenter une nouvelle méthode d'estimation de la tendance-cycle au moyen du filtre de Henderson à 13 termes ayant l'avantage de i) diminuer le nombre d'ondulations indésirables dans la courbe finale de la tendance-cycle, ii) réduire l'importance des corrections apportées aux valeurs concourantes préliminaires, et iii) ne pas augmenter le décalage de temps avec lequel sont décelés les points de retournement.

La nouvelle méthode consiste essentiellement à étendre la série désaisonnalisée lissée (modifiée par des valeurs extrêmes de poids nuls) d'une année au moyen d'extrapolations fondées sur le modèle ARMMI, puis, à appliquer le filtre de Henderson à 13 termes en fixant des limites standardisées strictes pour repérer et corriger les valeurs aberrantes.

L'application de la méthode à neuf séries tirées de l'Indice composite canadien des indicateurs avancés donne des résultats très satisfaisants.

REMERCIEMENTS

Le présent article se fonde sur des travaux financiers par Statistique Canada. Toutefois, les opinions exprimées sont celles de l'auteur et ne reflètent pas nécessairement celles de l'organisme. J'exprime ma gratitude à Marietta Morry et à Norma Chhab, du Centre de recherche et d'analyse en séries chronologiques, pour les innombrables discussions enrichissantes que nous avons eues et pour leur collaboration aux travaux qui ont mené à la publication de cet article. Je remercie également un évaluateur anonyme pour les commentaires précieux qu'il a faits au sujet d'une version antérieure.

BIBLIOGRAPHIE

CASTLES, I. (1987). A Guide to Smoothing Time Series Estimates of Trend. N° 1316.0 au catalogue, Australian Bureau of Statistics.

CHOLETTE, P.A. (1981). A comparison of various trend-cycle estimators. Dans *Time Series Analysis*. (O.D. Anderson et M.R. Perrymann, Eds). Amsterdam: North-Holland, 77-87.

CHOLETTE, P.A. (1982). Comparaison de deux estimateurs des cycles économiques. Document de recherche n° 82-09-OOIF, Centre de recherche et d'analyse en séries chronologiques, Statistique Canada.

CLEVELAND, R., CLEVELAND, W.S., McRAE, J.E., et TERPENNING, I. (1990). STL: A seasonal-trend decomposition procedure based on Loess. *Journal of Official Statistics*, 6, 3-33.

DAGUM, E.B. (1980). La méthode de désaisonnalisation X-11-ARMMI. N° 12-564F au catalogue, Statistique Canada.

DAGUM, E.B. (1988). The X-11-ARIMA/88 Seasonal Adjustment Method - Foundations and User's Manual. Centre de recherche et d'analyse en séries chronologiques, Statistique Canada.

DAGUM, E.B., et LANIET, N. (1987). Revisions of trend-cycle estimators of moving average seasonal adjustment methods. *Journal of Business and Economic Statistics*, 5, 177-189.

DAGUM, E.B., CHHAB, N., et CHIU, K. (1993). Linear properties of the X-11-ARIMA seasonal adjustment method. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*.

DAGUM, E.B., CHHAB, N., et CHIU, K. (1996). Derivation and properties of the Census X-11 variant and the X-11-ARIMA linear filters. *Journal of Official Statistics*, (à paraître).

FINDLEY, D.F., et MONSELL, B.C. (1990). Comment (sur Cleveland et coll. 1990). *Journal of Official Statistics*, 6, 55-59.

GRAY, A.G., et THOMSON, P.J. (1990). Comment (sur Cleveland et coll. 1990). *Journal of Official Statistics*, 6, 47-54.

HENDERSON, R. (1916). Note on graduation by adjusted average. *Transactions of the Actuarial Society of America*, 17, 43-48.

KENNY, P. (1993). Trend presentation. T02919, SMO, Branch, Central Statistical Office, Londres, Angleterre.

KENNY, P.B., et DURBIN, J. (1982). Local trend estimation and seasonal adjustment of economic and social time series. *Journal of the Royal Statistical Society, Series A*, 145, 1-41.

LESAGE, J.P. (1991). Analysis and development of leading indicators using a Bayesian turning-points approach. *Journal of Business and Economic Statistics*, 9, 305-316.

PFEFFERMANN, D., et BLEUER, S.R. (1992). Probabilistic detection of nonseasonal turning points in economic time series estimated from sample surveys. Rapport interne, Direction de la méthodologie, Statistique Canada, Ottawa.

RHOADES, D. (1980). La conversion de l'actualité en fiabilité des séries chronologiques économiques ou filtrage des séries chronologiques économiques occasionnant un déphasage minimum. *Revue statistique du Canada*, 6-13.

SCOTT, S. (1990). Comment (sur Cleveland et coll. 1990). *Journal of Official Statistics*, 6, 59-62.

SHISKIN, J., YOUNG, A.H., et MUSGRAVE, J.C. (1967). The X-11 Variant of Census Method II Seasonal Adjustment. Technical Paper No. 15, U.S. Bureau of the Census.

WALLGREN, B., et WALLGREN, A. (1990). Comment (sur Cleveland et coll. 1990). *Journal of Official Statistics*, 6, 39-46.

WECKER, W. (1979). Predicting the turning points of a series. *Journal of Business*, 52, 35-50.

ZELLNER, A., HONG, C., et MIN, C. (1991). Forecasting turning points in international output growth rates using Bayesian exponentially weighted autoregression, time-varying parameter, and pooling techniques. *Journal of Econometrics*, 48, 275-304.

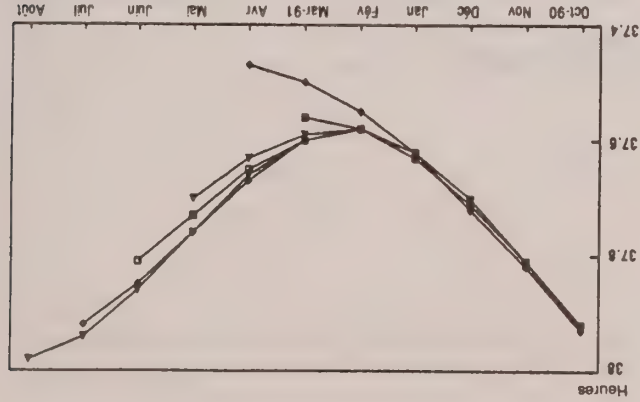


Figure 7b. Durée hebdomadaire de travail dans le secteur de fabrication. Evolution de la correction de l'estimation selon la méthode type de la tendance-cycle H13.

expliquées par les données sous-jacentes qui semblent signaler une baisse croissante contrédite par les valeurs suivantes. Les figures 7a, et 7b., qui correspondent à la Durée hebdomadaire du travail dans le secteur de la fabrication, indiquent que le point de retournement de février-mars 1991 est décalé avec un décalage de trois mois par les deux méthodes.

#### 4.3 Diminution de la correction des estimations concurrentes de la tendance-cycle

La réduction de la correction totale des estimations les plus récentes de la tendance-cycle, qui sont de nature provisoire, est un autre aspect important dont il faut tenir compte. Théoriquement, on obtient l'estimation finale de la tendance-cycle après avoir ajouté quatre années de données à la série, mais les corrections deviennent très petites après l'ajout des données de trois mois.

Le tableau 2 montre le pourcentage moyen absolu de correction des estimations concurrentes de la tendance-cycle durant une période de quatre ans allant de janvier 1988 à décembre 1991. Les résultats indiquent que, dans six des neuf cas analysés, la correction totale des valeurs concurrentes de la tendance-cycle est beaucoup plus faible pour la méthode modifiée que pour la méthode type, et qu'elle n'est légèrement plus importante que dans deux cas.

Tableau 2

Pourcentage moyen absolu de correction totale des valeurs concurrentes de la tendance-cycle au moyen du filtre de Henderson à 13 termes

Séries	Méthode type (1)	Méthode modifiée (2)	Rapport (2)/(1)
NCBD	1.55	1.10	0.73
VDMAM	0.62	0.47	0.76
VDBD	0.77	0.62	0.80
RLS	0.87	0.70	0.80
DHTF	0.13	0.12	0.92
TSE300	1.12	1.07	0.95
M1	0.35	0.35	1.00
IL	2.09	2.20	1.05
ESPE	0.40	0.42	1.05

Figure 6a. Nouvelles commandes de biens durables. Evolution de la correction de l'estimation selon la méthode modifiée de la tendance-cycle H13.

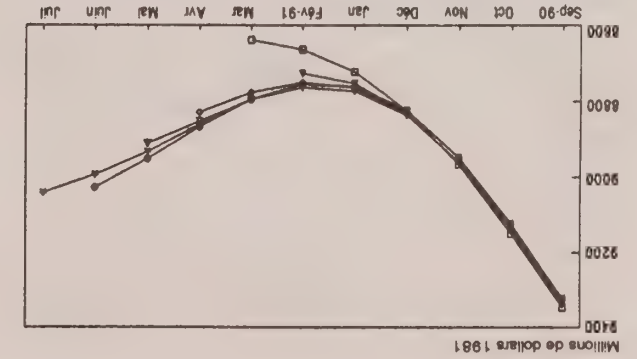


Figure 6b. Nouvelles commandes de biens durables. Evolution de la correction de l'estimation selon la méthode type de la tendance-cycle H13.

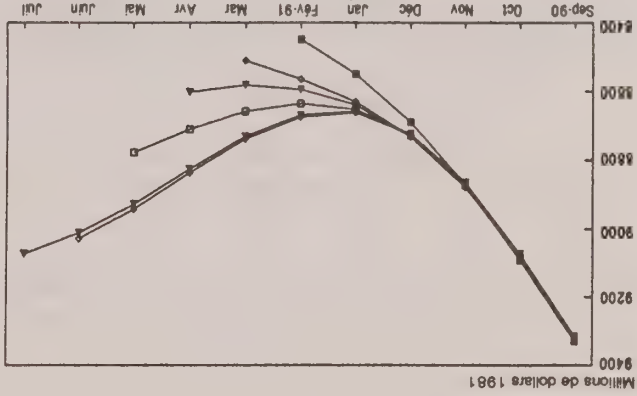
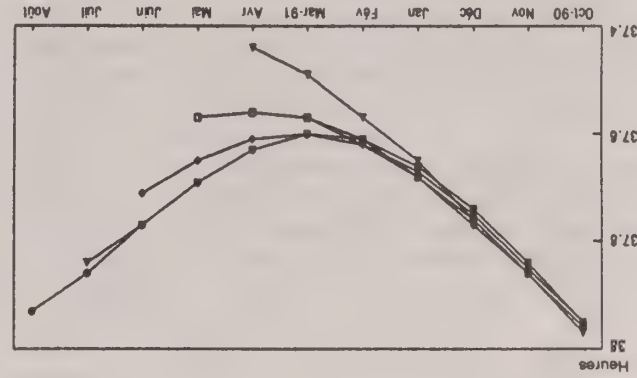
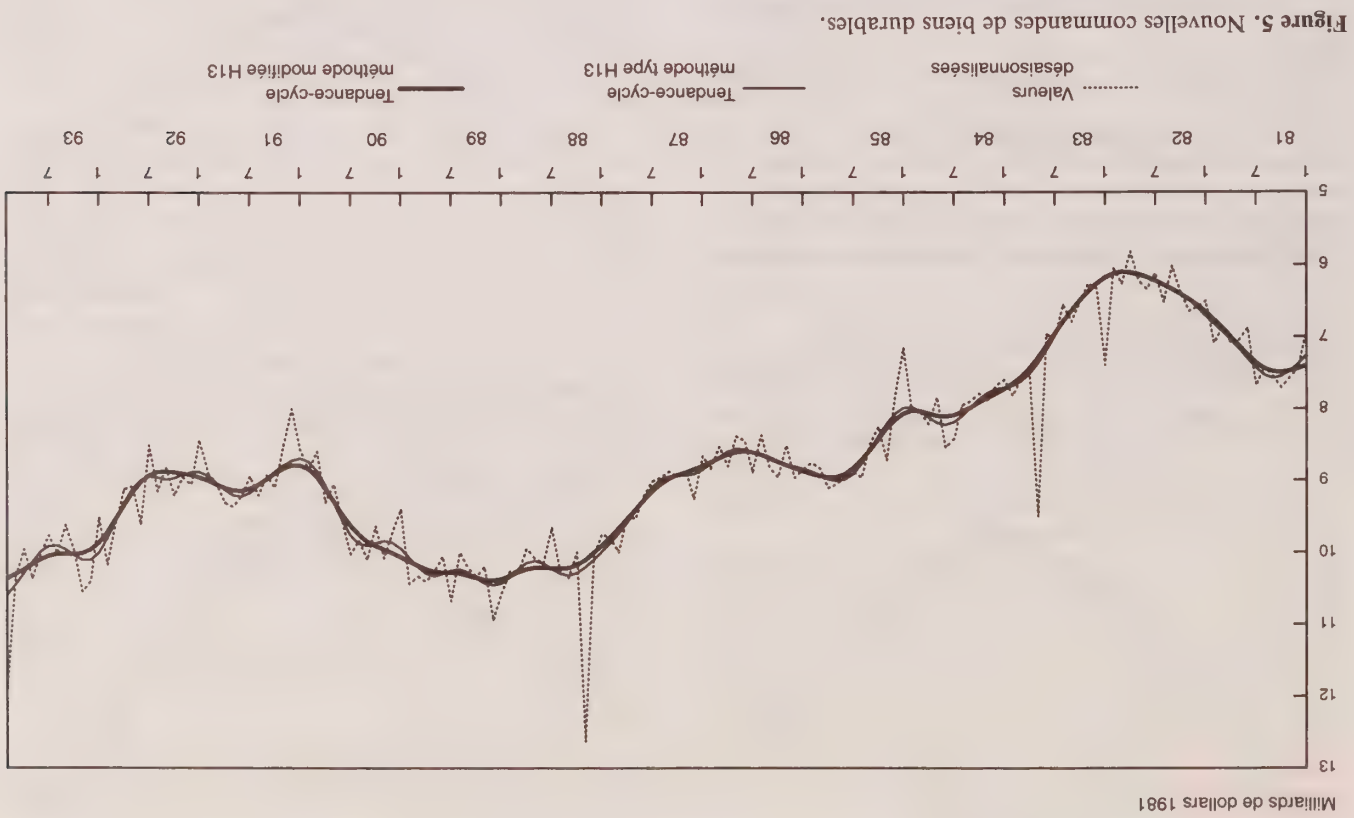


Figure 7a. Durée hebdomadaire de travail dans le secteur de fabrication. Evolution de la correction de l'estimation selon la méthode modifiée de la tendance-cycle H13.



et ainsi de suite. Le point de retournement est reconnu en avril, après deux mois, tandis qu'en appliquant la méthode type, sa détection prend trois mois, comme l'illustre la figure 6b. En outre, les figures 6a, et 7a, montrent que les estimations révisées successives de la tendance-cycle restent généralement très proches des valeurs finales. Les lignes qui font saillie, indiquant une correction importante, sont





Le tableau 1 montre le nombre d'ondulations observé dans les estimations de la tendance-cycle quand on applique le filtre type ou le filtre de Henderson à 13 termes modifié pour la période de janvier 1981 à décembre 1993.

**Tableau 1**  
Nombre d'ondulations indésirables dans les données sur la tendance-cycle obtenues en utilisant le filtre de Henderson à 13 termes pour la période de 1981 à 1993

Séries	Méthode type	Méthode modifiée
NCBD	9	2
IL	8	4
VDBD	8	4
ESPE	8	5
DHTF	7	1
RLS	5	1
TSE300	4	2
MI	4	2
VDMAM	4	0

Les résultats indiquent que, d'une part, la réduction du nombre d'ondulations est d'autant plus importante que ce nombre est élevé au départ et, d'autre part, qu'elle est très significative dans tous les cas.

À titre d'illustration, les figures 4 et 5, correspondant aux séries DHTF et NCBD respectivement, montrent les valeurs désaisonnalisées et les données sur la tendance-cycle obtenues en appliquant les méthodes type et modifiée. Ces figures montrent que la nouvelle méthode permet de

réduire le nombre d'ondulations décelées dans la courbe de la tendance-cycle, comparativement à la méthode type. En fait, la tendance-cycle obtenue selon la méthode modifiée ressemble à celle produite au moyen du filtre de Henderson à 23 termes, quoiqu'elle coupe davantage les crêtes et les creux des cycles de longue période soit plus prononcée.

#### 4.2 Détection des points de retournement

Il importe que la réduction du nombre d'ondulations dans les estimations finales de la tendance-cycle n'augmente pas le décalage avec lequel sont décelés les points de retournement, phénomène qui constitue la principale limitation du filtre de Henderson à 23 termes.

Pour examiner l'évolution de la correction de la tendance-cycle à n'importe quel point dans le temps, on a calculé les estimations pour tous les points finals et pour les points précédents dans le temps. La comparaison de l'évolution de la correction de l'estimation de la tendance-cycle selon la méthode modifiée et selon la méthode type montre qu'en moyenne, le décalage avec lequel les points de retournement du cycle sont décelés est similaire pour les deux méthodes. Selon la série examinée, l'écart observé entre les décalages est nul, ou de l'ordre de plus ou moins un mois. À titre d'illustration, la figure 6a, montre l'évolution de la correction de l'estimation, selon la méthode modifiée, de la tendance-cycle des Nouvelles commandes de biens durables pour le point de retournement du cycle de février 1991. Des mises à jour successives sont effectuées au moyen de données allant jusqu'à mars 1991, avril 1991

retournelement comme un point  $t$  dans le temps où une des observations d'une série, disons  $Y_t$ , a une valeur plus élevée (ou plus faible) que les  $k$  observations précédentes et les  $m$  observations subséquentes, ou est égale à celles-ci. Autrement dit,

$$Y_{t-k} \leq \dots \leq Y_{t-1} > Y_t \geq Y_{t+1} \geq \dots \geq Y_{t+m}$$

définit le passage à une phase descendante, et

$$Y_{t-k} \geq \dots \geq Y_{t-1} < Y_t \leq Y_{t+1} \leq \dots \leq Y_{t+m}$$

le passage à une phase ascendante.

En ce qui concerne les séries désaisonnalisées ou les données sur la tendance-cyclo, il n'existe pas de consensus quant aux valeurs de  $k$  et de  $m$  qui définissent un point de retournelement. Rhoades (1980) considère qu'un point de

retournelement est atteint quand  $k = 1$  et  $m = 0$ ; Wecker (1979) définit un point de retournelement comme la deuxième d'au moins deux baisses ou hausses successives, c.-à-d.  $k = 2$  et  $m = 2$ , et d'autres, dont Zellner, Hong et Min (1991), LeSage (1991), et Pfeiffermann et Bleuer (1992) choisissent  $k = 3$  et  $m = 0$ . Ces définitions des points de retournelement ne correspondent pas nécessairement à celles utilisées pour l'analyse des cycles économiques, mais n'importe laquelle permet de calculer le nombre d'ondulations, à condition que deux points de retournelement (un passage à la baisse et un passage à la hausse) surviennent durant une période de 10 mois ou moins. Nous considérons ici qu'un point de retournelement survient quand  $k = 3$  et  $m = 0$ , car les données sur la tendance-cyclo sont déjà très lisses.

nouvelle méthode sort du cadre de la présente étude, mais il est vraisemblable que les seconds seraient également meilleurs (consulter Cholette 1982). La plupart des séries choisies sont extrêmement volatiles et mènent toutes à des points de retournelement du cycle économique. Les séries étudiées sont les suivantes:

Indice du cours des actions TSE300  
Indice du logement (IL)  
Offre de monnaie (M1)  
Emploi dans les services aux personnes et aux entreprises (ESPE)  
Durée hebdomadaire du travail dans le secteur de fabrication (DHTF)  
Ventes au détail de meubles et d'articles ménagers (VDMAM)  
Ventes au détail d'autres biens durables (VDBD)  
Nouvelles commandes de biens durables (NCBD)  
Ratio des livraisons aux stocks (RLS).

Les avantages de la nouvelle méthode comparativement à la méthode X-11-ARMMI utilisée actuellement sont évalués en regard des points suivants.

#### 4.1 Réduction du nombre d'ondulations dans les estimations finales de la tendance-cyclo

Afin de calculer la réduction du nombre d'ondulations, nous commençons par donner la définition d'un point de retournelement dans le contexte des données de la tendance-cyclo. On définit ordinairement un point de

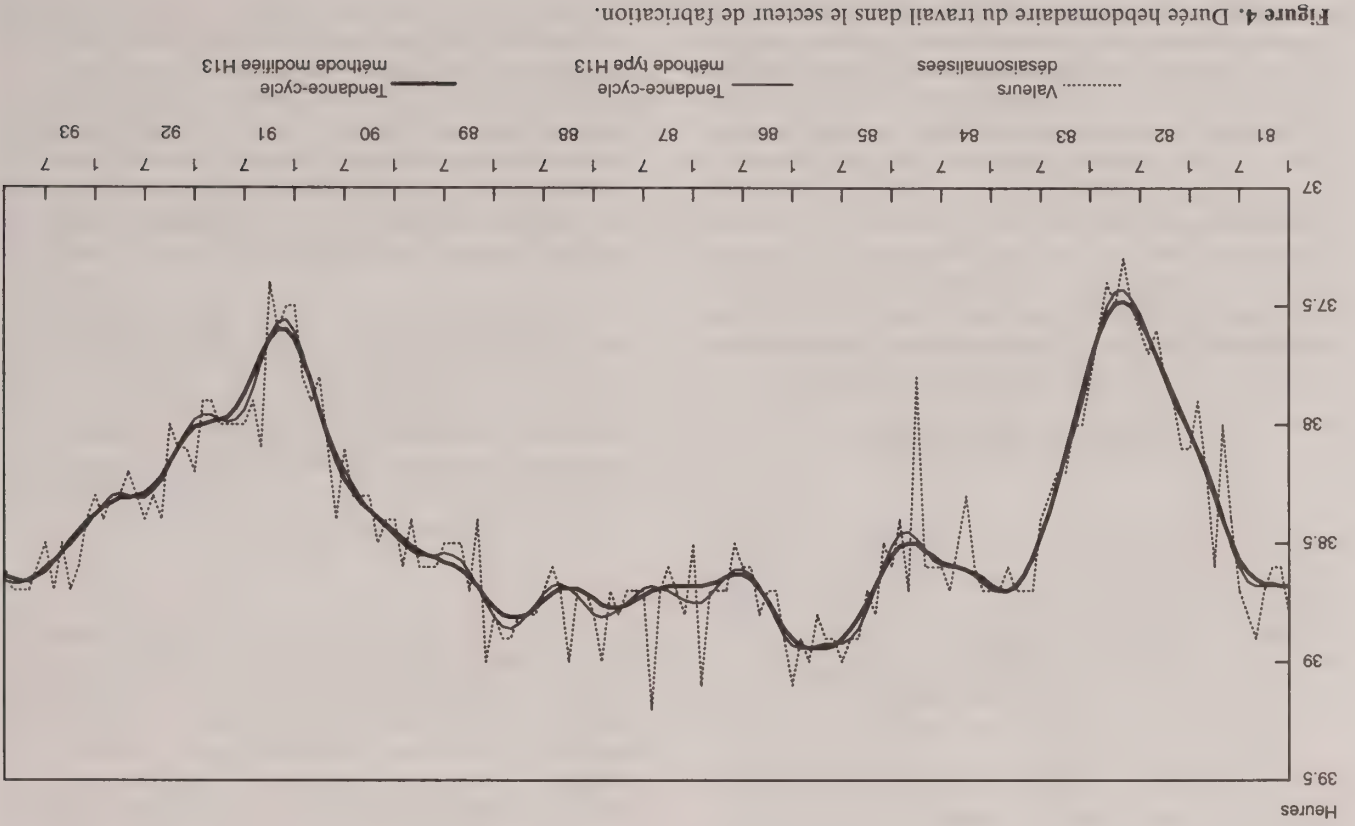


Figure 4. Durée hebdomadaire du travail dans le secteur de fabrication.



Le gain et le déphasage obtenus avec ce filtre concourant de la tendance-cycle, qui ne sont pas illustrés faute d'espace, se situent entre ceux produits par les deux autres combinaisons.

Dans le cas d'extrapolations ARMMI, le gain produit par le filtre concourant converge rapidement vers la valeur finale. Dagum et Laniel (1987) montrent qu'après l'ajout de trois observations supplémentaires à la série, le gain obtenu avec le filtre asymétrique est très proche de celui obtenu avec le filtre symétrique. Les propriétés de ces filtres de la tendance-cycle sont également examinées en détail par Dagum, Chhab et Chiu (1993 et 1996).

Le filtre de Henderson à 13 termes produit des ondulations dans l'estimation finale de la tendance-cycle unique-ment si les données entrées dans le filtre, c.-à-d. la série désaisonnalisée dont on a remplacé les valeurs extrêmes, présentent une certaine amplitude pour la bande de fréquences de 0.08 à 0.16.

Dans la plupart des cas empiriques, la présence d'ondulations indésirables s'observe pour des périodes de grande volatilité durant lesquelles les données observées sont influencées surtout par des valeurs aberrantes qui peuvent être interprétées erronément comme des points de retournement du cycle. Bien que les séries désaisonnalisées soient corrigées pour tenir compte des valeurs extrêmes, il est nécessaire de les lisser davantage, soit en appliquant un filtre de Henderson plus long, soit en effectuant plus strictement la correction des valeurs extrêmes. Afin de bénéficier de la plus grande rapidité de détection des points de retournement qu'offre un filtre court, la seconde méthode est celle choisie ici.

À l'heure actuelle, les limites implicites sur lesquelles on se fonde pour corriger les valeurs extrêmes qui influent sur la série désaisonnalisée sont fixées à  $\pm 1.5$  sigma et  $\pm 2.5$  sigma. On attribue un poids nul aux valeurs situées au-delà de  $\pm 2.5$  sigma et un poids égal à un à celles situées en deçà de  $\pm 1.5$  sigma. Aux valeurs irrégulières comprises entre  $\pm 2.5$  sigma et  $\pm 1.5$  sigma, on attribue un poids tractionnaire qui varie linéairement de zéro à un.

### 3. NOUVELLE MÉTHODE

La nouvelle méthode proposée consiste essentiellement 1) à étendre la série désaisonnalisée lissée (modifiée par des valeurs extrêmes ayant un poids nul) au moyen d'extrapolations ARMMI, et 2) à appliquer le filtre de Henderson à 13 termes à la série étendue en fixant des limites standardisées plus strictes pour repérer et corriger les valeurs extrêmes.

L'expérimentation avec des données réelles montre que le spectre d'amplitude de la série désaisonnalisée n'est réduit radicalement aux fréquences variant de 0.8 à 0.16 que si on fixe des limites standardisées strictes, telles que  $\pm 0.7$  sigma et  $\pm 1.0$  sigma. Dans ces conditions, la courbe de la tendance-cycle obtenue quand on applique le filtre de Henderson à 13 termes ne présente pas d'ondulations indésirables, mais permet toujours de déceler rapidement les points de retournement. Si on émet l'hypothèse que la

distribution est normale, ces nouvelles limites standardisées impliquent que 48% des valeurs irrégulières seront modifiées, 32% recevront un poids nul, et 16%, un poids tractionnaire variant de zéro à un.

L'extension de la série désaisonnalisée lissée au moyen d'extrapolations ARMMI est nécessaire pour diminuer la correction des estimations les plus récentes de la tendance-cycle. Si on se sert du modèle X-1-ARMMI ou X11, l'application de la nouvelle méthode doit comporter les deux étapes suivantes:

1) La meilleure série désaisonnalisée possible est produite en choisissant les options les plus appropriées pour l'estimation des composantes, c'est à dire la saisonnalité, la tendance-cycle, les variations des jours commerciaux et les effets de Pâques, plus les valeurs antérieures permanentes et temporaires, le cas échéant. Les valeurs désaisonnalisées sont présentées au tableau D11. La série désaisonnalisée, corrigée pour tenir compte des valeurs extrêmes de poids nul repérées conformément aux limites standardisées implicites, figure au tableau E2. Si on modifie les estimations de la série désaisonnalisée publiée pour l'année courante conformément à des méthodes de correction particulières, il faut traiter de nouveau la série révisée au moyen du programme X-11-ARMMI pour obtenir les données de sortie correspondantes au tableau E2.

2) La série de données de sortie du tableau E2 est étendue d'une année grâce à des données extrapolées au moyen d'un modèle ARMMI. Le modèle (0,1,1) (0,0,1) se révèle adéquat pour nombre de séries réelles. Bien que les données du tableau E2 soient désaisonnalisées, l'emploi du paramètre saisonnier à moyennes mobiles (dont la valeur est souvent très faible) est nécessaire pour tenir compte de certaines autocorrélations saisonnières. On traite ensuite la série étendue au moyen du programme X-11-ARMMI, en choisissant l'option de mesures sommaires et en appliquant des limites standardisées strictes ( $\pm 0.7$  sigma et  $\pm 1.0$  sigma) ainsi que le filtre de Henderson à 13 termes. Les estimations de la tendance-cycle ainsi obtenues figurent au tableau D12.

### 4. RÉSULTATS EMPIRIQUES

La nouvelle méthode d'estimation de la tendance-cycle est testée au moyen de neuf séries tirées de l'Indice composite canadien des indicateurs avancés. Dans la version dite «filtrée» de l'Indice composite des indicateurs avancés publiée par Statistique Canada, on lisse chaque série consécutivement, ainsi que l'Indice, en appliquant aux données désaisonnalisées des filtres asymétriques fondés sur le modèle ARMMI élaboré par Rhoades (1980). Les propriétés spectrales de ces filtres ARMMI de la tendance-cycle sont similaires à celles des points extrêmes des filtres de Henderson à 9, à 13 ou à 23 termes, selon le modèle ARMMI choisi. La comparaison des résultats obtenus avec les filtres ARMMI à ceux observés en appliquant la

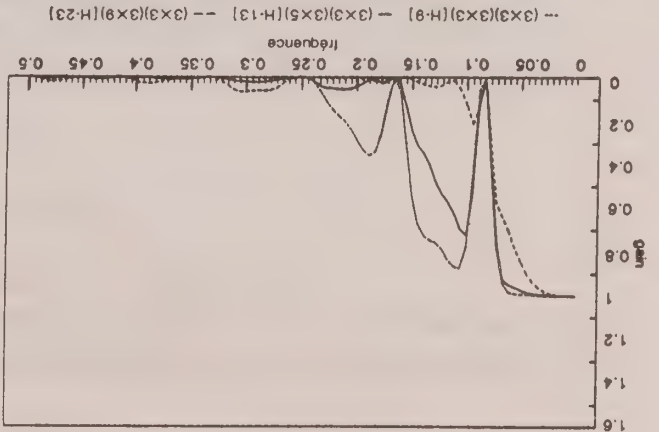


Figure 1. Filtrés en cascade symétriques pour l'estimation de la tendance-cycle.

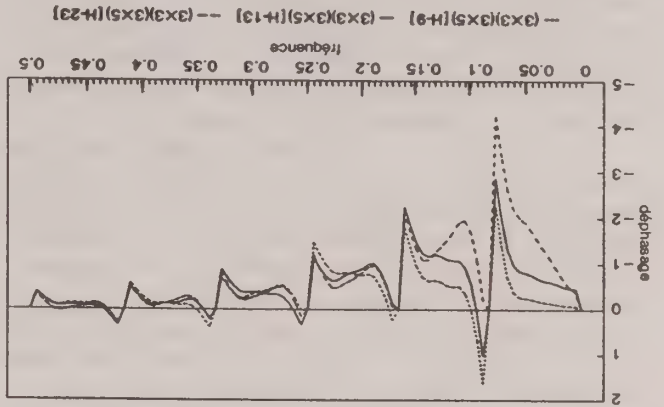
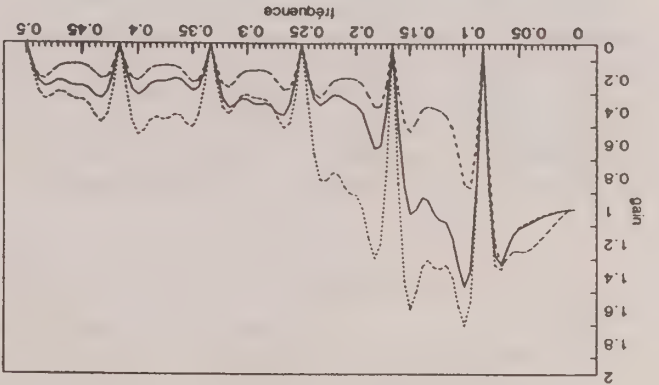


Figure 2. Filtrés en cascade concourants de la tendance-cycle. Moyenne mobile saisonnière type combiné aux trois filtres de Henderson.

cascade symétrique applicable aux valeurs finales ou centrales (au moins quatre années à partir de chaque extrémité de la série) produisant une amélioration de l'ordre de celle illustrée à la figure 1.

La figure 1 montre aussi les améliorations obtenues avec d'autres convolutions de filtres, nommément 1) des filtres saisonniers courts combinés au filtre de Henderson

à 9 termes et 2) des filtres saisonniers longs combinés au

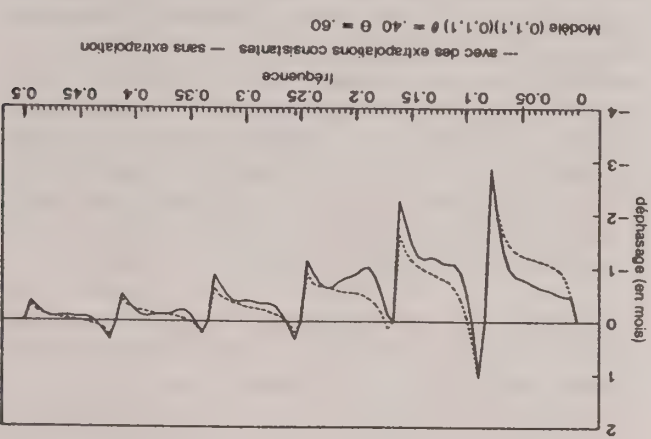
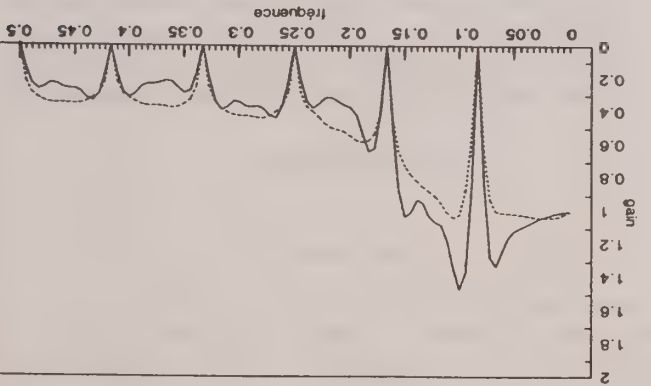


Figure 3. Filtrés en cascade concourants de la tendance-cycle, ARMMI.  $(3 \times 3) (3 \times 5) [H - 13]$ , avec et sans extrapolations

La figure 3 montre que, quand on utilise des extrapolations ARMMI, les gains obtenus au moyen du filtre en cascade concourant (contenant le filtre de Henderson à 13 termes) se rapprochent de ceux produits par les filtres symétriques, moyennant toutefois une légère augmentation du déphasage. Les extrapolations sont celles d'un modèle ARMMI  $(0,1,1,1) (0,1,1)$ , où le paramètre ordinaire à moyennes mobiles est  $\theta = 0,40$  et le paramètre saisonnier à moyennes mobiles est  $\Theta = 0,60$ .

Si on applique les filtres concourants de la tendance-cycle aux observations les plus récentes, on observe un sommet encore plus élevé dans la bande de fréquences qui correspond aux cycles de 9 et de 10 mois (figure 2). En outre, ces filtres asymétriques produisent tous un déphasage, pratiquement de deux mois pour le filtre à 23 termes, d'un mois pour celui à 13 termes et d'un demi-mois pour celui à 9 termes.

l'absence de la tendance-cycle, aucun des filtres en cascade utilisés pour estimer la tendance-cycle, particulièrement ceux contenant le filtre de Henderson à 9 ou à 13 termes, ne supprime les cycles de 9 ou 10 mois (bande de fréquences de 0,08 à 0,16). En fait, le filtre en cascade symétrique obtenu au moyen du filtre de Henderson à 9 termes laisse passer environ 90% de l'amplitude de ces cycles courts; ceux contenant le filtre à 13 ou à 23 termes laissent passer 72% et 21%, respectivement.



# Nouvelle méthode visant à limiter le nombre d'ondulations indésirables et les corrections lors de l'estimation de la tendance-cycle au moyen du modèle X-11-ARMMI

ESTELA BEE DAGUM<sup>1</sup>

## RÉSUMÉ

L'estimation de la tendance-cycle par la méthode X-11-ARMMI est souvent effectuée en appliquant le filtre de Henderson à 13 termes à des données désaisonnalisées, modifiées par des valeurs extrêmes. Cependant, ce filtre produit dans la courbe tendance-cycle finale ou «historique» un grand nombre d'ondulations indésirables qui sont interprétées, erronément, comme des points de retournement. L'utilisation d'un filtre de Henderson plus long, tel que celui à 23 termes, n'est pas une solution, car ce filtre retarde la détection des points de retournement, donc, ne convient pas pour les analyses économiques et commerciales courantes. L'auteur propose une nouvelle méthode d'utilisation du filtre de Henderson à 13 termes ayant l'avantage de i) diminuer le nombre d'ondulations indésirables, ii) réduire l'importance des corrections apportées aux valeurs préliminaires et iii) ne pas augmenter le décalage avec lequel sont décelés les points de retournement. Les résultats de l'application de la méthode à neuf indicateurs avancés de l'Indice composite canadien des indicateurs avancés sont présentés à titre d'illustration.

MOTS CLÉS: Tendance-cycle; X-11-ARMMI; points de retournement; indicateurs économiques avancés.

## 1. INTRODUCTION

L'estimation de la tendance-cycle par la méthode de désaisonnalisation X-11-ARMMI (Dagum 1980, 1988), ainsi que par la variante X-11 du U.S. Bureau of the Census (Shiskin, Young et Musgrave 1967), est effectuée au moyen de filtres linéaires élaborés par Henderson (1916). Les filtres sont appliqués à des séries désaisonnalisées dont les données irrégulières sont modifiées pour tenir compte des valeurs extrêmes. La longueur du filtre est sélectionnée automatiquement, d'après des valeurs spécifiques du ratio signal/bruit (I/S), le filtre à 13 termes étant le plus couramment.

La question de l'estimation de la tendance-cycle a été étudiée par plusieurs auteurs, dont Rhoades (1980), Cholette (1981, 1982), Kenny et Durbin (1982), Castles (1987), Dagum et Lanjel (1987), Cleveland, McKrae et Terpenning (1990), Wallgren et Wallgren (1990), Gray et Thomson (1990), Findley et Monsell (1990), Scott (1990), et Kenny (1993). Cependant, la plupart des bureaux de la statistique (sauf l'Australian Bureau of Statistics) se concentrent avant tout sur la publication de données désaisonnalisées, quelques-uns seulement fournissant des renseignements sur la tendance-cycle, ordinairement sous forme graphique. Plusieurs raisons poussent les bureaux de la statistique à limiter la publication d'estimations de la tendance-cycle. En général, les séries désaisonnalisées sont déjà suffisamment lisses pour donner une indication claire de la tendance à court terme. Dans le cas des séries très volatiles, qui exigent un lissage supplémentaire, les principales objections à l'estimation de la tendance-cycle sont: 1) l'ampleur de la correction des valeurs les plus récentes (en général, beaucoup plus importante que pour les estimations

désaisonnalisées correspondantes) et 2) la présence de cycles courts ou ondulations (cycles de 9 et 10 mois) dans la courbe finale de la tendance-cycle quand on utilise le filtre de Henderson à 13 termes. À cet égard, Kenny (1993) argumente que la présence d'ondulations dans les estimations finales de la tendance-cycle provoque la détection d'un grand nombre de faux points de retournement, phénomène qui rend le filtre à 13 termes inapproprié pour la surveillance des renversements de cycle. Il propose l'utilisation du filtre de Henderson à 23 termes en vue d'obtenir une courbe beaucoup plus lisse. Cependant, il est bien connu que ce filtre plus long retarde la détection des points de retournement, donc, ne convient pas aux analyses économiques et commerciales courantes. Pour ce type d'analyse, le filtre de Henderson à 13 termes est préférable. Toutefois, il présente l'inconvénient de produire des ondulations susceptibles d'être interprétées erronément comme des points de retournement. La présente étude vise principalement à proposer une méthode d'utilisation du filtre de Henderson à 13 termes ayant l'avantage de 1) réduire le nombre d'ondulations indésirables, 2) limiter l'ampleur des corrections apportées aux estimations les plus récentes quand de nouvelles observations sont ajoutées à la série et 3) ne pas augmenter le décalage avec lequel sont décelés les points de retournement.

## 2. FILTRES EN CASCADE POUR L'ESTIMATION DE LA TENDANCE-CYCLE

Le filtre de Henderson à 13 termes est celui choisi le plus fréquemment. Combiné aux filtres saisonniers type (moyennes mobiles à 5 et à 7 termes), il donne un filtre en

<sup>1</sup> Estela Bee Dagum, Faculté des sciences statistiques, Université de Bologne, Via delle Belle Arti 41, (40126) Bologne, Italie.





REMERCIEMENTS

Les auteurs tiennent à remercier M. Michel Hidiroglou pour ses commentaires utiles. Nous remercions également Carol (Veum) Caldwell, Easley Hoy, les critiques de la revue Techniques d'enquête et les gestionnaires du service de la recherche et de la méthodologie de la Manufacturing and Construction Division pour leurs commentaires utiles formulés au cours de l'examen du présent article.

BIBLIOGRAPHIE

COCHRAN, W.G. (1977). *Sampling Techniques*, (3<sup>e</sup> éd.). New York: John Wiley and Sons.

DETLEFSEN, R., et VEBUM, C. (1991). Design issues for the retail trade sample surveys of the U.S. Bureau of the Census. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 214-219.

ECKMAN, G. (1959). An approximation useful in univariate stratification. *The Annals of Mathematical Statistics*, 30, 219-229.

HESS, I., SETHI, V.K., et BALAKRISHNAN, T.R. (1966). Stratification: A practical investigation. *Journal of the American Statistical Association*, 61, 74-90.

HIDIROGLOU, M.A. (1986). The construction of a self-representing stratum of large units in survey design. *The American Statistician*, 40, 27-31.

LAVALLÉE, P., et HIDIROGLOU, M.A. (1988). Sur la stratification de populations asymétriques. *Techniques d'enquête*, 14, 35-45.

SCHNEEBERGER, H. (1979). Saddle-points of the variance of the sample mean in stratified sampling. *Sankhya, Series C*, 41, 92-96.

SETHI, V.K. (1963). A note on optimum stratification of populations for estimating the population means. *American Journal of Statistics*, 5, 20-23.

1) la différence entre la nouvelle borne supérieure et la borne supérieure de l'itération précédente est inférieure à un. Nous utilisons le nombre entier de un, dans notre cas, puisque les valeurs de la paye n'existent qu'en nombres entiers et qu'un changement des bornes d'une valeur inférieure à un n'influera sur aucune entreprise; 2) la différence entre la nouvelle borne inférieure et celle de l'itération précédente est inférieure à un; 3) la différence entre la nouvelle taille de l'échantillon et celle de l'itération antérieure est inférieure à une petite valeur arbitraire. Nous recommandons un nombre inférieure à un puisque les valeurs de la taille de l'échantillon sont habituellement arrondies et que toute amélioration fractionnaire de la taille de l'échantillon est négligeable. Il convient d'être prudent lors du choix de cette valeur puisque'il est possible que le taux de réduction de la taille de l'échantillon augmente d'une itération à l'autre par suite d'un changement de la pente de la surface; 4) le programme atteint la 30<sup>e</sup> itération. Il s'agit évidemment la d'une valeur arbitraire qui peut dépendre du nombre de fois (industries) où il convient d'utiliser la méthode L-H.

Il convient en outre de mentionner que les populations petites peuvent provoquer une convergence des bornes à un point sous-optimal, comme nous l'avons montré dans les exemples. Les graphiques de la taille de l'échantillon montrent une surface rugueuse pour les petites populations et une surface lisse pour les grandes. C'est cette surface rugueuse, due à la nature discrète de la petite population, qui contribue en partie à déterminer le lieu de convergence de la méthode L-H.

Souignons finalement, à titre de conclusion, que la méthode de Dalenius-Hodges présume que toutes les strates obtenues seront échantillonnées. La méthode L-H est conçue pour établir une sous-strate analytique à tirage complet. Par conséquent, la strate supérieure élaborée par la méthode de Dalenius-Hodges au moment de la création des bornes initiales pour les branches ACES, sera déséquilibrée puisqu'elle ne sera pas échantillonnée. Dans une telle situation, nous avons observé des améliorations dans la taille de l'échantillon en passant de la méthode de Dalenius-Hodges à la première itération de la méthode L-H. Il peut cependant arriver que les bornes de départ conduisent à un minimum local qui ne représente pas la meilleure solution.

les bornes finales du tableau 3 ainsi que d'autres bornes avec leurs valeurs correspondantes de la taille de l'échantillon. Il semble que la méthode L-H converge vers une région inférieure sur l'une des crêtes principales, à condition que cette région se trouve dans le voisinage des bornes optimales. La taille d'échantillon minimale est de 9.22 et de 9.36. La valeur entière la plus petite de la taille de l'échantillon pour chaque résultat qui répond à la condition ou qui la dépasse est 10. Ici encore, nous constatons que la méthode L-H fonctionne exceptionnellement bien même avec des distributions discrètes dont les populations sont petites, puisque les bornes convergent à l'intérieur du voisinage contenant la solution optimale. Il convient en outre de signaler que les bornes peuvent prendre une large gamme de valeurs tout en maintenant la même valeur entière de la taille de l'échantillon. À mesure que la taille du voisinage s'élargit, la gamme des bornes augmente également. Signalons finalement que même si la gamme des valeurs  $b_1$  pour un voisinage donné est plus petite que la gamme des valeurs  $b_2$ , il existe beaucoup plus d'unités de sondage dans la gamme des  $b_1$  que dans celle des  $b_2$  à cause de l'asymétrie de la distribution.

## 5. CONCLUSION

Les graphiques que nous avons présentés montrent qu'une vaste gamme de bornes donne une gamme étroite de valeurs de la taille de l'échantillon lorsqu'elle se trouve au voisinage d'une valeur optimale (portion inférieure en forme de cuvette des graphiques). Une amélioration additionnelle de la taille de l'échantillon (p. ex., un léger gain marginal) risque de ne pas justifier les efforts supplémentaires nécessaires pour l'obtenir. Il n'est même pas sûr qu'un tel gain marginal contribuera à améliorer la taille de l'échantillon puisque cette dernière est en réalité une valeur entière et que le gain marginal en question pourrait n'être qu'une petite fraction. La méthode L-H s'est avérée très efficace pour le calcul des bornes dans un voisinage souhaité autour d'une valeur optimale, et elle s'est également avérée relativement rapide. En mesurant le taux de convergence à l'aide de la taille de l'échantillon au lieu des bornes, nous avons été en mesure de mieux déterminer le moment où on atteignait un voisinage souhaitable autour d'une valeur optimale. Les bornes varient en effet grandement dans un tel voisinage tandis que la taille de l'échantillon (le paramètre qui nous intéresse surtout) ne varie que légèrement. L'absence de l'amélioration obtenue dans la taille de l'échantillon d'une itération à l'autre était devenue minimale ou nulle, nous arrêtons immédiatement le programme en pressurant que nous avons atteint le voisinage désiré. Les règles d'inter-ruption suivantes sont recommandées. On interrompt le programme lorsque:

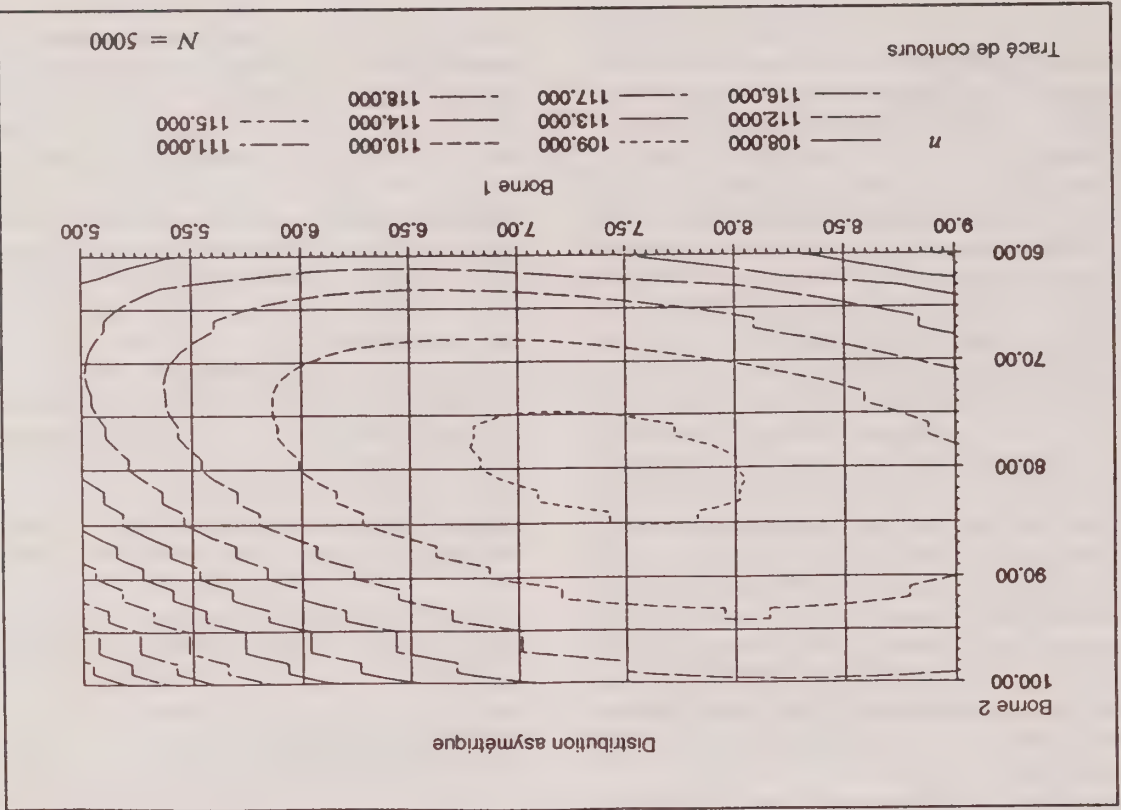
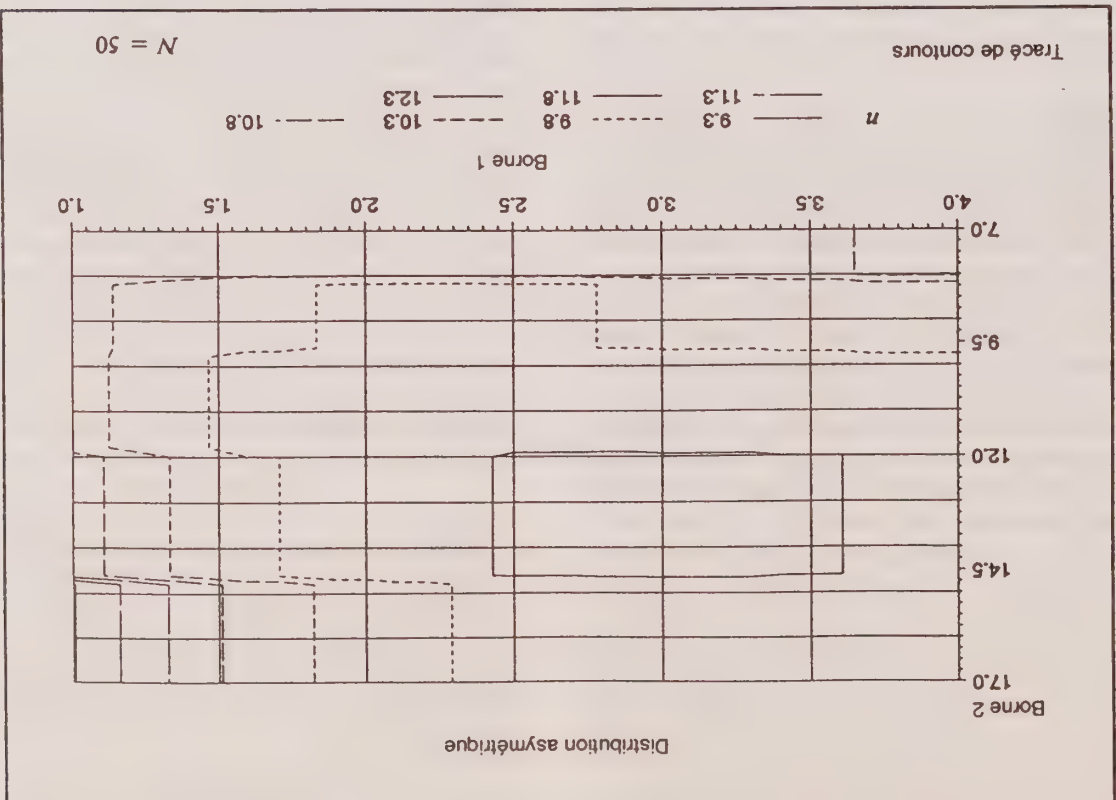
Nous avons procédé ainsi pour maintenir l'extrémité supérieure de la distribution discrète finie proportionnellement comparable, en gros, à la population entière pour chaque taille de population. À cette fin, nous avons choisi le paramètre  $b$  de manière qu'une proportion d'environ 90% de la somme totale soit prise en compte par les 20% supérieurs de l'ensemble des unités de sondage possibles. Puisque les ensembles de données contiennent un nombre fini de valeurs discrètes, la dérivation des variances des strates différentes lorsque les valeurs de  $b$  étaient inférieures à 2 n'a pas posé de problème.

Nous présentons au tableau 3 les résultats de l'application de la méthode L-H pour des tailles de population et des points de départ différents. Le premier groupe utilise des valeurs de départ qui donnent des populations de strates égales ( $N_1 = N_2 = N_3$ ). Le deuxième groupe utilise la méthode de Dalenius-Hodges pour obtenir l'ensemble des bornes initiales. Le troisième groupe obtient les bornes initiales en utilisant une méthode de détermination de la borne de la strate à tirage complet proposée par Hidiroglou (1986) et utilise la méthode de Dalenius-Hodges pour les autres bornes. Ici encore, on peut observer que la surface représentant la taille de l'échantillon est beaucoup plus rugueuse pour les petites populations (voir figure 5). Par exemple, lorsque  $N = 50$  et que  $b_1$  est fixe, on n'obtient qu'une seule taille de l'échantillon lorsque  $b_2$  varie entre 11.8 et 14.7. Ceci s'explique par l'absence de valeurs comprises à l'intérieur de cette gamme dans la population. À mesure que la taille de la population augmente, les valeurs des données se rapprochent et la surface de l'échantillon s'adoucit (voir figure 6).

Le tracé de contours correspondant à  $N = 50$  (figure 7) présente des formes irrégulières délimitées par des lignes droites. Par contre, le tracé de contours correspondant à  $N = 5000$  (figure 8) présente des formes elliptiques concentriques presque lisses. Une forme comparable et concentrique des tracés de contours semblerait être une qualité souhaitable puisque cela signifierait que le minimum global est le seul minimum local.

Le tracé de contours correspondant à  $N = 50$  illustre le cas où la méthode L-H n'a pas donné une convergence vers des bornes optimales. Puisque, dans cet exemple, nous avons laissé le programme L-H tourner jusqu'à la convergence, il est permis de se demander pourquoi la méthode L-H n'a pas abouti à une convergence aux bornes optimales. Ce problème s'explique plus facilement à l'examen de la figure 5. Nous constatons en effet que lorsque la taille de la population est petite, la surface de la taille de l'échantillon n'est pas aussi lisse que dans la figure 6. Nous distinguons dans la figure 5 des crêtes importantes causées par les vastes lacunes dans les données discrètes asymétriques ( $x_3 = 9.71, x_4 = 11.81, x_5 = 14.79, x_6 = 19.29$ ). Ceci signifie que pour une valeur donnée  $b_1$ , toute valeur de  $b_2$  comprise entre 11.81 et 14.79 donnerait la même taille d'échantillon. Lorsque nous avons utilisé le programme L-H avec des bornes de départ différentes au tableau 3, nous avons obtenu que les trois énumérées au tableau 3, nous avons obtenu



Figure 8. Tracé de contours d'une distribution asymétrique ( $N = 5000$ ).Figure 7. Tracé de contours d'une distribution asymétrique ( $N = 50$ ).

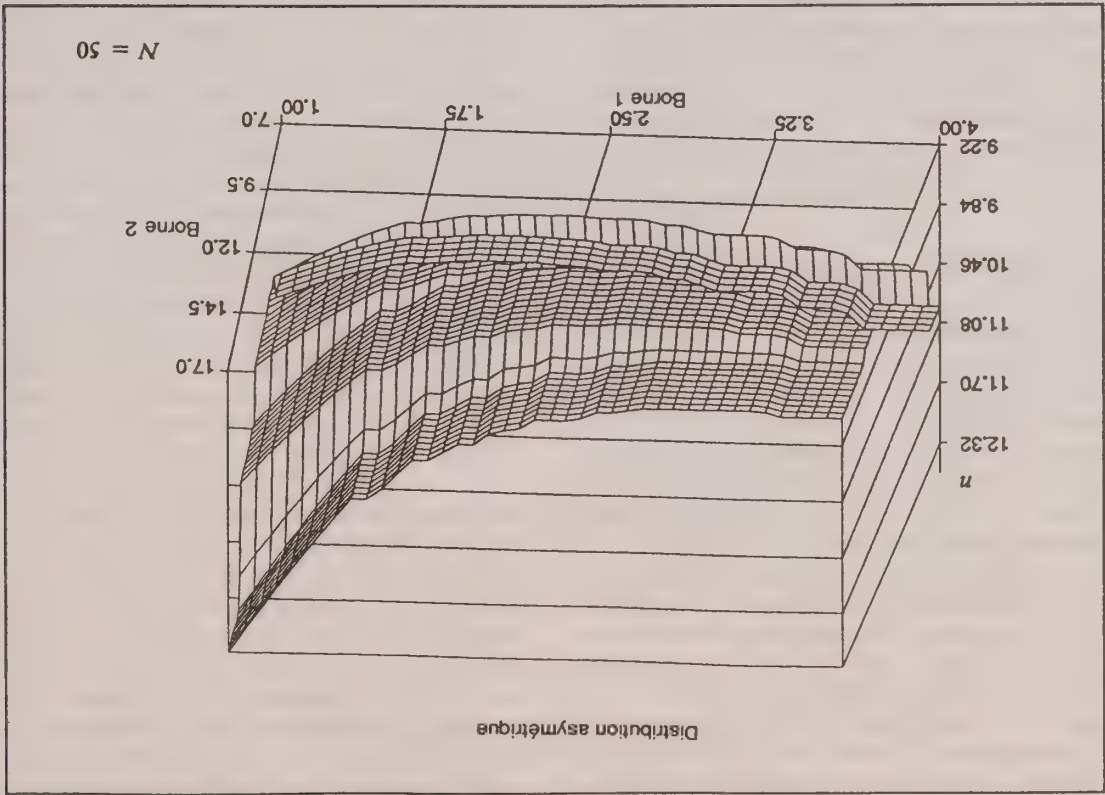


Figure 5. Surface de la taille de l'échantillon pour une distribution asymétrique ( $N = 50$ ).

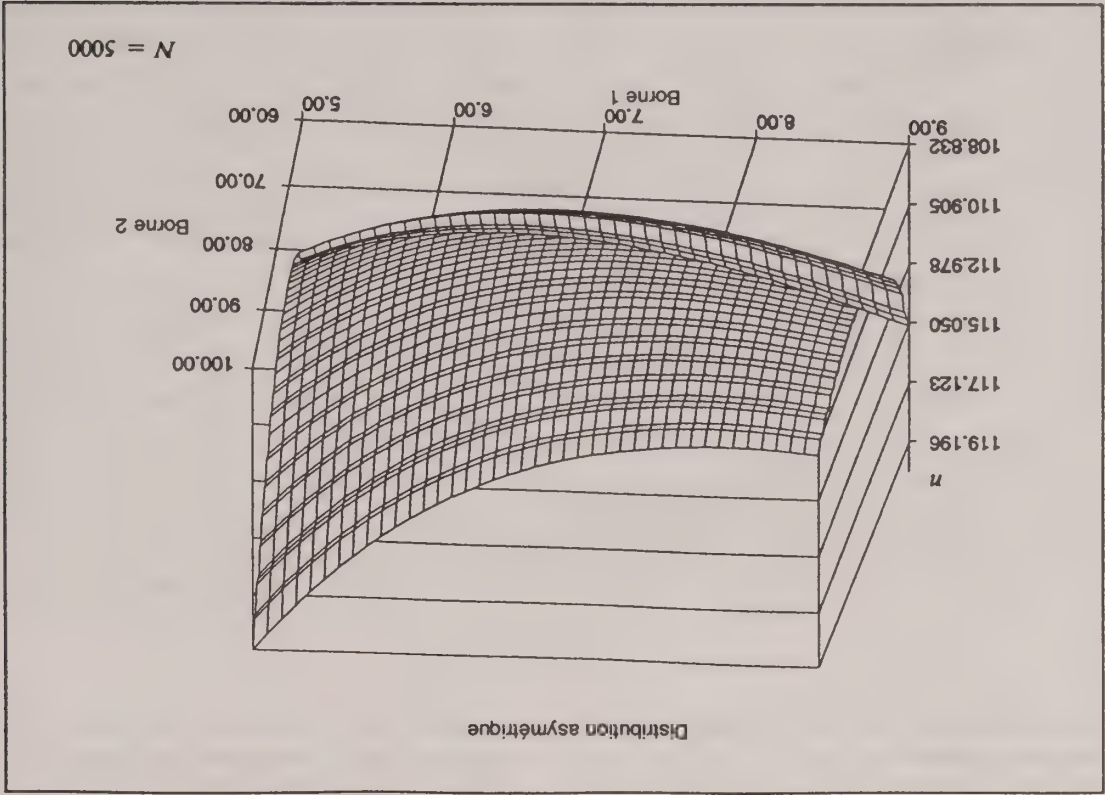


Figure 6. Surface de la taille de l'échantillon pour une distribution asymétrique ( $N = 5000$ ).



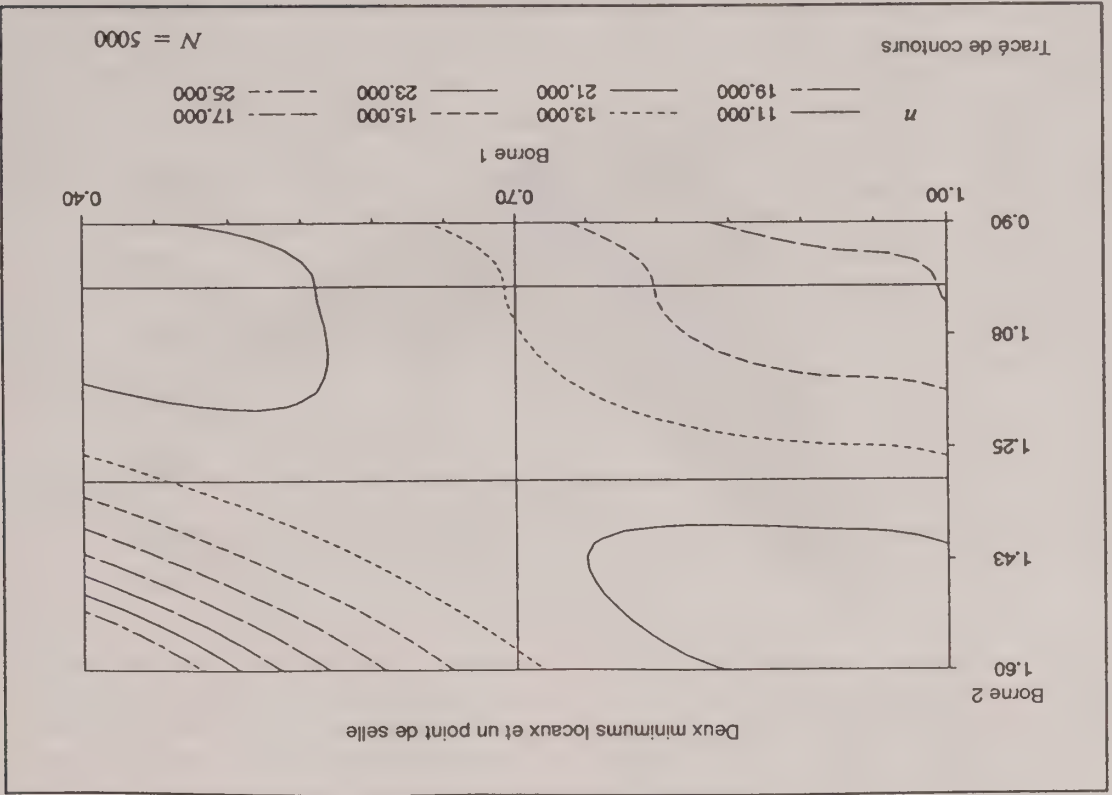
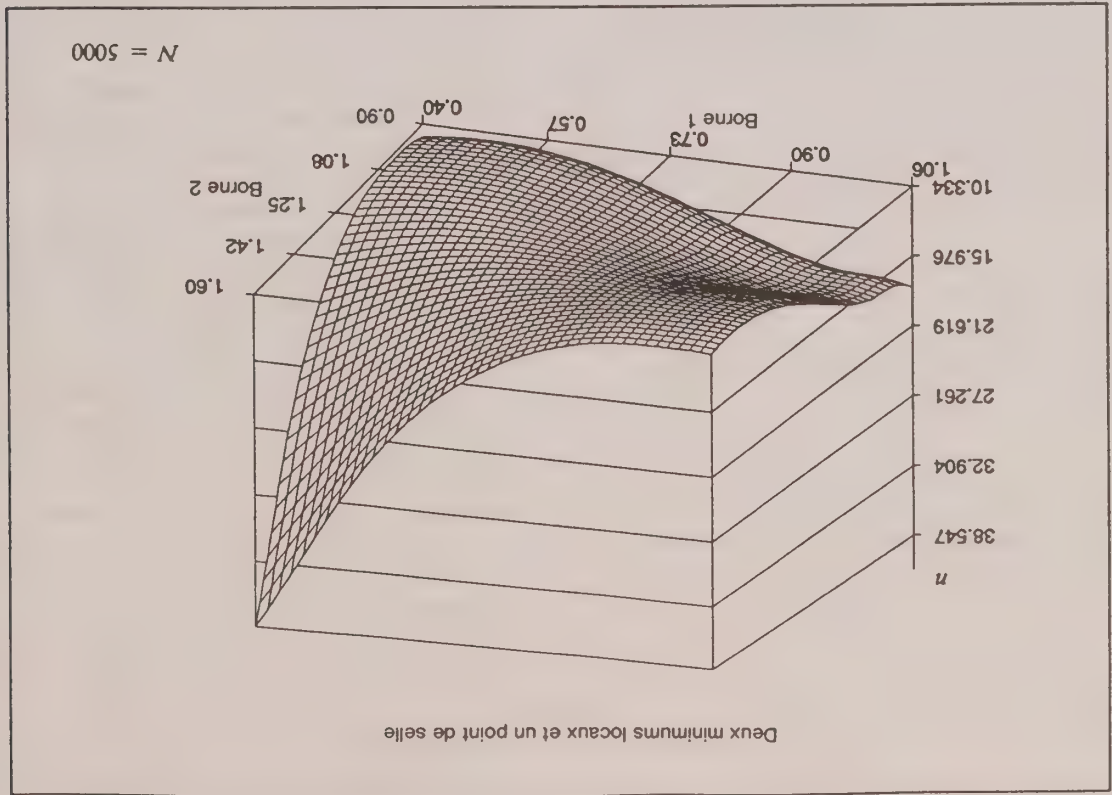
la méthode de Dalenius-Hodges, lesquelles ne se situent pas sur la ligne lorsque  $b_2 > 2 - b_1$ , la méthode L-H converge vers un minimum (2c). La méthode de Dalenius-Hodges fonctionne bien dans cet exemple à cause des trois strates à tirage partiel. Avec des bornes de départ qui ne se situent pas sur la ligne dans le cas où  $b_2 < 2 - b_1$  (dans ce cas-ci,  $b_1 = .5$  et  $b_2 = 1.3$ ), la méthode L-H converge vers un minimum différent (2a). Ce problème n'est pas unique à la méthode L-H puisque Schneeberger indique que les bornes obtenues par la méthode du gradient sont également dépendantes des bornes de départ.

On observe finalement dans cet exemple que les quelques premières itérations donnent des réductions relativement grandes de la taille de l'échantillon et que les itérations qui suivent donnent plusieurs réductions plus petites de cette taille. Nous présentons au tableau 2 les résultats de l'itération dans laquelle l'algorithme produit une taille de l'échantillon qui s'écarte de moins de 5% de la taille finale. Ceci signifie que l'algorithme L-H tend rapidement vers un voisinage autour de la borne optimale. Lorsque nous sommes près d'une taille optimale de l'échantillon, il semble exister une vaste gamme de bornes produisant un nombre limité de tailles d'échantillon. L'application de règles d'interruption peut effectivement réduire le temps de calcul sans nous conduire à renoncer à une véritable réduction de la taille de l'échantillon puisque les valeurs de ce paramètre sont données en nombres entiers.

Nous présentons à la figure 4 un tracé de contours de la surface illustrée à la figure 3. Ici encore, les axes représentent les bornes inférieures et supérieures des strates et la surface est définie par la taille d'échantillon qui en découle. Les lignes du tracé représentent une valeur de taille de l'échantillon. L'espace délimité correspond à une gamme de valeurs de la taille de l'échantillon. Par exemple, la ligne continue

Bornes L-H d'une distribution asymétrique (une strate à tirage complet, deux strates à tirage partiel)

N	Méthode de démarrage	1ère itération					Itération à moins de 5% de la taille de l'échantillon					Itération finale				
		b	b <sub>1</sub>	b <sub>2</sub>	nTA	n	b	b <sub>1</sub>	b <sub>2</sub>	nTA	n	b	b <sub>1</sub>	b <sub>2</sub>	nTA	n
50	$N_1 = N_2 = N_3$	.80	.63	2.81	17	17.2	1.66	10.20	7	9.6	5	.80	2.44	11.81	7	9.4
100	$N_1 = N_2 = N_3$	.90	.56	2.33	34	34.3	1.61	10.29	11	15.8	5	.90	2.58	12.44	10	15.1
200	$N_1 = N_2 = N_3$	.90	.56	2.36	67	67.2	2.35	17.04	15	21.8	6	.90	3.61	20.46	13	20.9
1000	$N_1 = N_2 = N_3$	1.00	.50	2.00	333	334.2	3.35	30.58	32	53.0	7	1.00	4.93	36.32	27	51.3
5000	$N_1 = N_2 = N_3$	1.05	.47	1.85	1665	1667.2	4.67	64.33	62	113.5	7	1.05	7.39	79.38	50	108.8
50	Dalenius-Hodges	.80	1.25	8.04	9	10.5	1.76	10.37	7	9.5	3	.80	2.44	11.81	7	9.4
100	Dalenius-Hodges	.90	1.39	8.98	13	16.6	1.62	10.16	11	15.8	2	.90	2.58	12.44	10	15.1
200	Dalenius-Hodges	.90	1.82	11.66	20	24.3	2.45	17.29	15	21.7	3	.90	3.61	20.46	13	20.9
1000	Dalenius-Hodges	1.00	2.37	17.28	55	65.6	3.15	29.70	33	53.5	3	1.00	4.93	36.32	27	51.3
5000	Dalenius-Hodges	1.05	3.09	26.27	155	175.0	4.98	66.28	60	112.3	4	1.05	7.39	79.38	50	108.8
50	Hiditgloou 1986	.80	.94	6.50	10	11.3	1.58	10.02	7	9.6	3	.80	2.44	11.81	7	9.4
100	Hiditgloou 1986	.90	.74	6.17	17	19.6	1.66	10.38	11	15.8	4	.90	2.58	12.44	10	15.1
200	Hiditgloou 1986	.90	1.39	9.55	24	27.2	2.50	17.58	14	21.5	4	.90	3.61	20.46	13	20.9
1000	Hiditgloou 1986	1.00	2.02	15.13	62	71.3	3.34	30.54	32	53.0	4	1.00	4.93	36.32	27	51.3
5000	Hiditgloou 1986	1.05	3.24	28.72	142	164.1	5.11	67.05	59	112.0	4	1.05	7.39	79.38	50	108.8





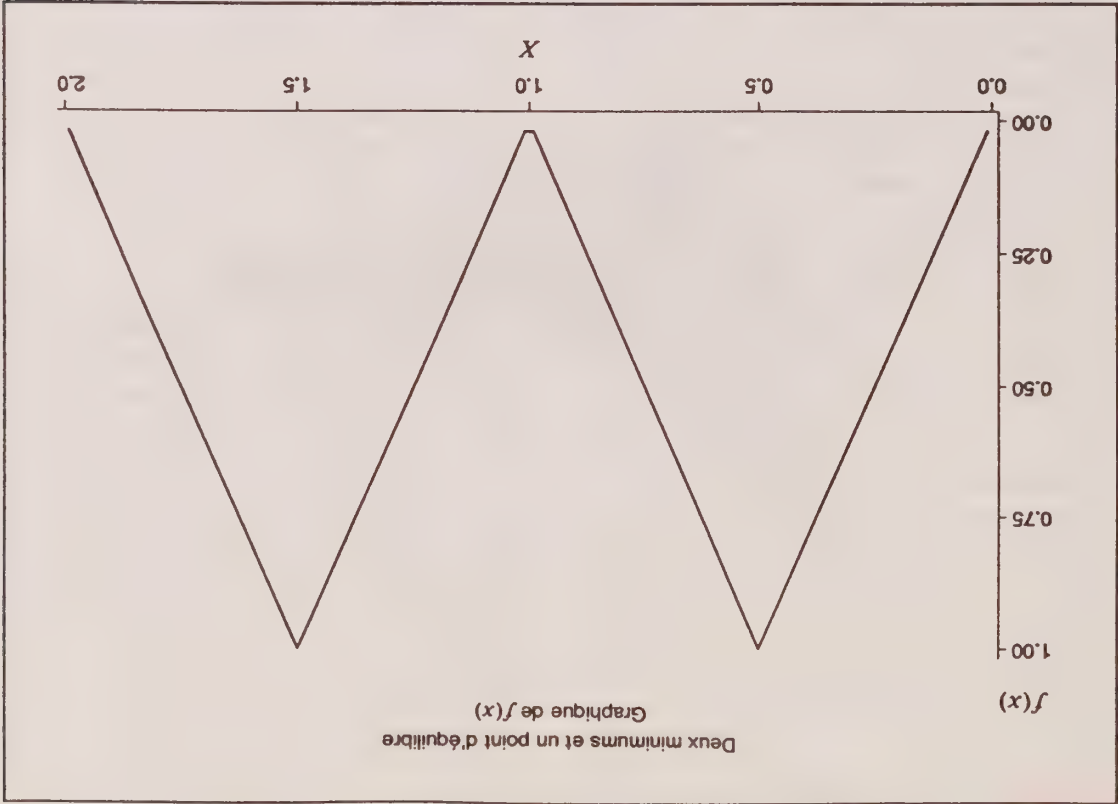


Figure 1. Graphique d'une distribution symétrique.

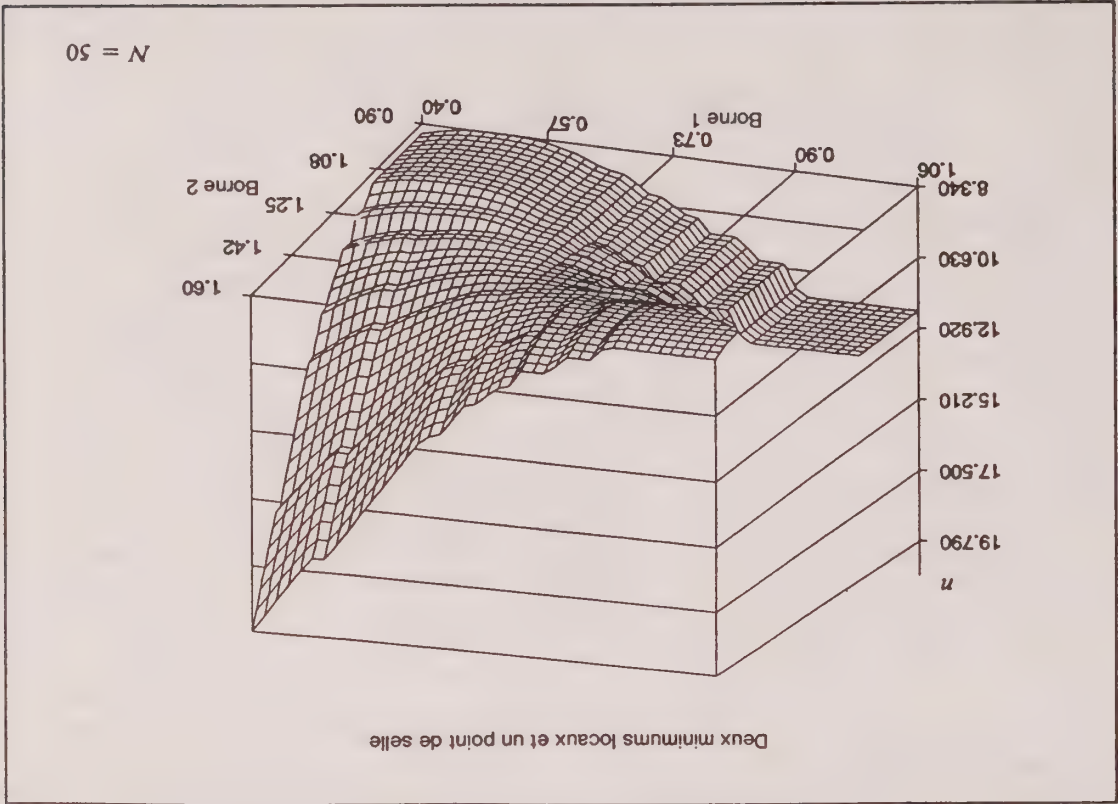


Figure 2. Surface de la taille de l'échantillon pour une distribution symétrique ( $N = 50$ ).

Schneeberger cherchait à définir les bornes de trois strates à tirage partiel en utilisant une méthode du gradient. Les résultats obtenus à l'aide de la fonction objective de  $z = (\sum W_h o_h)^2$  sont présentés au tableau 1.

**Tableau 1**  
Bornes optimales pour une distribution non asymétrique

	$b_1$	$b_2$	Point optimal
(2a)	.50241	1.03985	Point minimal
(2b)	.70910	1.29090	Point de selle
(2c)	.96015	1.49759	Point minimal

Source: Schneeberger (1979).

Nous avons généré cinq ensembles de données de tailles différentes ( $N = 50, 100, 200, 1,000$  et  $5,000$ ) en utilisant la formule  $F(x) = (j - 1/2)/N$ . Pour cet exemple, nous avons adapté la méthode L-H afin d'établir trois strates à tirage partiel et aucune à tirage complet et de pouvoir ainsi comparer nos résultats avec ceux de l'article de Schneeberger. Dans notre application de l'estimation des totaux, avec minimisation de la taille de l'échantillon sous réserve d'une valeur du c.v. de 5%, nous utilisons la méthode L-H pour chacune des cinq tailles de population en utilisant trois méthodes différentes de démarrage. Les résultats de cette opération sont présentés au tableau 2.

Cet exemple nous permet en outre de rappeler encore une fois que les bornes finales dépendent des bornes initiales. Pour cet exemple, Schneeberger indique qu'avec un point de départ symétrique à  $x = 1$ , où  $b_1 = 1 - \lambda$  et  $b_2 = 1 + \lambda$  ( $0 < \lambda < 1$ ) et qui définit la ligne  $b_2 = 2 - b_1$ , la méthode du gradient déplace le gradient sur la ligne  $b_2 = 2 - b_1$  jusqu'au point d'équilibre. Lorsque nous posons les bornes de départ sur cette ligne, comme cela s'est produit lorsque nous avons commencé avec  $N_1 = N_2 = N_3$ , la méthode L-H converge également au point d'équilibre (voir tableau 1). Avec les bornes de départ de

**Tableau 2**  
Bornes L-H de trois strates à tirage partiel pour une distribution symétrique

N	Méthode de démarrage	1 <sup>re</sup> itération			Itération à moins de 5% de la taille de l'échantillon			Itération finale		
		$b_1$	$b_2$	$n$	$b_1$	$b_2$	$n$	$b_1$	$b_2$	itér. n°
50	$N_1 = N_2 = N_3$	.59	1.41	10.89	.66	1.34	9.98	.70	1.31	2
100	$N_1 = N_2 = N_3$	.59	1.41	12.60	.66	1.34	10.91	.70	1.30	2
200	$N_1 = N_2 = N_3$	.59	1.41	13.42	.66	1.34	11.43	.71	1.29	2
5000	$N_1 = N_2 = N_3$	.59	1.41	13.85	.66	1.34	11.75	.71	1.29	2
50	Dalenius-Hodges	.70	1.40	10.09	.70	1.40	10.09	.77	1.37	1
100	Dalenius-Hodges	.70	1.40	10.90	.84	1.40	10.14	.93	1.47	7
200	Dalenius-Hodges	.70	1.40	11.42	.83	1.40	10.44	.95	1.49	7
5000	Dalenius-Hodges	.70	1.40	11.86	.86	1.42	10.67	.96	1.50	8
50	Hors ligne	.50	1.30	10.87	.57	1.20	9.43	.55	1.11	3
100	Hors ligne	.50	1.30	11.95	.57	1.18	10.04	.53	1.07	3
200	Hors ligne	.50	1.30	12.64	.56	1.14	10.28	.51	1.05	4
5000	Hors ligne	.50	1.30	13.37	.56	1.14	10.67	.50	1.04	4
50	Hors ligne	.50	1.30	10.87	.57	1.20	9.43	.55	1.11	3
100	Hors ligne	.50	1.30	11.95	.57	1.18	10.04	.53	1.07	3
200	Hors ligne	.50	1.30	12.64	.56	1.14	10.28	.51	1.05	4
5000	Hors ligne	.50	1.30	13.37	.56	1.14	10.67	.50	1.04	4



où  $n_{TA}$  désigne le nombre d'entreprises dans la strate à tirage complet à l'intérieur de la strate II définie par la méthode L-H,  $N$  désigne le nombre d'entreprises de la strate II dans la branche ACES qui nous intéresse,  $W_j = N_j/N$  désigne la proportion de la strate,  $N_j$  désigne

$$n = n_{TA} + \frac{\frac{cv^2 \chi^2}{2} + \sum_{j=1}^f \frac{N}{W_j S_j^2}}{N \left( \sum_{j=1}^f \frac{W_j S_j}{2} \right)^2} \quad (1)$$

L'application de la méthode L-H au plan de sondage préliminaire de l'ACES de 1992 exigeait la séparation de la strate II en une strate à tirage complet et en deux strates à tirage partiel pour chaque industrie de l'ACES. Les bornes ont été dérivées pour chaque industrie en prenant la dérivée partielle de la taille de l'échantillon pour une des bornes, tout en fixant l'autre. Toutefois en pratique, nous avons permis aux deux bornes de se déplacer simultanément. Nous avons ainsi obtenu un processus itératif de minimisation de la taille de l'échantillon pour chaque industrie, compte tenu des contraintes imposées par le c.v. À l'intérieur de la strate II, pour chaque industrie de l'ACES, et en tenant pour acquise la répartition de Neyman (Delforsen et Veum 1991), l'équation de la taille de l'échantillon qui est minimisée se présente comme suit:

s'amenuiseront. élevée, les possibilités de répondre aux exigences de fiabilité comme la corrélation entre la paye et les dépenses n'est pas de ce qui est réellement nécessaire. En conséquence, il se peut que les tailles d'échantillons soient différentes, la taille de l'échantillon est directement liée à la variance, variation de la paye et la variation des dépenses. Comme On s'inquiète par ailleurs d'un écart possible entre la emprise sur la fiabilité des estimations.

dans laquelle elles ont été classées, on perd alors toute signalent des dépenses dans des branches autres que celle dépenses dans de multiples branches. Si trop d'entreprises dans le questionnaire ACES. Elles peuvent indiquer des leurs dépenses en capital consacrées aux branches ACES élevée. Toutefois, les entreprises indiquent elles-mêmes strate II dans la branche où elle versait la paye la plus taille choisie. Nous avons classé chaque entreprise de la classification des entreprises par suite de la mesure de la Ce plan de sondage présente un risque de mauvaise annuelle de l'entreprise à l'aide de la méthode L-H.

sur la taille ont été créées en fonction de la paye totale chaque catégorie d'industrie ACES, trois strates fondées classée. Par la suite, à l'intérieur de la strate II, pour totale estimée versée dans la branche dans laquelle elle était d'une entreprise donnée contrairement à la paye dans lesquelles elle était active. Toutefois, la feuille de paye informations de base pour chacune des branches ACES à plus d'une activité. Chaque entreprise possédait des classées en une seule industrie, même si elles s'adonnaient 500 employés. Les entreprises de la strate II ont été La strate II contenait les entreprises comptant entre cinq

Notre premier exemple est une distribution symétrique tirée de l'article de Schneebberger. Cette distribution est symétrique au point  $x = 1$ , comme l'illustre la figure 1.

#### 4.1 Distribution symétrique

Il peut parfois arriver que la méthode L-H exige un grand nombre d'itérations avant que les bornes ne convergent; il arrive même que ces bornes ne convergent jamais. En règle générale, après quelques itérations seulement, une vaste proportion des améliorations possibles de la taille de l'échantillon ont déjà été obtenues. Notre but était de mettre en pratique des règles d'interruption qui permettraient d'arrêter le processus dès qu'on atteindrait une certaine superficie autour d'un minimum local. C'est ce qui nous a poussés à utiliser des traces de contours dans notre analyse de l'effet des bornes sur la taille de l'échantillon obtenu. Cela nous a également permis d'obtenir un aperçu graphique des voisinages des minimum locaux. Nous utiliserons deux distributions pour illustrer les avantages d'une révision des traces de contours.

#### 4. CONVERGENCE EN VOISINAGES

Le taux de fiabilité pour chaque industrie correspondait à une valeur attendue du c.v. de 5% sur la paye. Nous ignorons toutefois quelles seraient les valeurs de l'écart-type découlant des dépenses en capital, puisqu'il n'existait pas de données sur les dépenses en capital pour la base de sondage. Les entreprises répondant dans des branches ACES différentes de celles pour lesquelles elles avaient contribué dans le plan de sondage ont également causé des fluctuations des valeurs du c.v. Le nombre total d'entreprises sélectionnées pour l'enquête préliminaire de l'ACES de 1992 était de 11,194, soit 1,500 entreprises de la strate I et 9,694 entreprises de la strate II.

où  $y_{ji}$  désigne la valeur de la paye de l'entreprise  $i$  appartenant à la strate  $j$  pour la branche ACES qui nous intéresse et  $\bar{Y}_j$  désigne la moyenne de la paye pour la strate  $j$ .

$$S_j = \sqrt{\frac{\sum_{i=1}^{N_j} (y_{ji} - \bar{Y}_j)^2}{N_j - 1}}$$

$N_j$  désigne le nombre d'entreprises de la strate I et  $S_j$  désigne l'écart-type de la paye tiré de la liste SSBL pour la strate  $j$  dans la strate II définie par

$$Y = \sum_{N_I}^k y_k + \sum_{N_j}^3 y_{ji},$$

définie par strates I et II de la branche ACES qui nous intéresse, désigne le coefficient de variation désiré pour la branche le nombre d'entreprises de la strate II pour la strate  $j$ , c.v.



du gradient, la solution obtenue peut être une valeur minimale, maximale, tant locale que globale, ou encore un point de selle de la variance de la moyenne de l'échantillon. Detlefsen et Veum (1991) ont jugé que cette caractéristique de la méthode L-H constituait un inconvénient lorsqu'ils en ont étudié l'application pour l'enquête mensuelle sur le commerce au détail du Bureau of the Census. Ils ont observé que les bornes obtenues avec la méthode L-H étaient souvent sensiblement différentes des bornes fixées au départ, de sorte que la taille minimale de l'échantillon obtenu correspondait à un minimum local. Représentée géométriquement, la taille de l'échantillon en fonction de deux bornes de stratification prend l'aspect d'un paysage marqué d'une ou plusieurs vallées en forme de cuvettes. La méthode L-H commence dans une région et descend jusqu'au point le plus bas. S'il existe plus d'un minimum, elle ne poursuit pas sa recherche du minimum global. En conséquence, un des objectifs visés consiste à avoir des bornes initiales qui se situent au voisinage du minimum global. Le recours à des bornes de départ découlant d'une méthode comme celle de Dalenius et Hodges pourrait permettre de répondre à cette exigence.

Detlefsen et Veum (1991) ont également relevé des cas de convergence lente ou de non-convergence. Toutefois, ils ont également noté que la convergence est plus rapide lorsque le nombre de strates est réduit et lorsque les bornes de départ sont les mêmes que les bornes de sélection de l'échantillon de l'enquête précédente. Afin de nous prémunir contre un nombre de boucles infinies dans le programme informatique ou contre un nombre très élevé d'itérations, nous avons décidé de faire deux choses: premièrement, nous avons mis en oeuvre un plan d'échantillonnage dans lequel la méthode L-H créerait des ensembles de seulement trois strates fondées sur la taille. Deuxièmement, nous avons décidé de mettre en oeuvre des règles d'interruption qui feraient en sorte d'interrompre le programme lorsque le taux de convergence semblait diminuer.

Dans le présent article, nous fournissons des informations de base sur l'ACES et nous décrivons brièvement notre utilisation de la méthode L-H. Nous montrons comment les tracés de contours et les graphiques tridimensionnels justifient l'utilisation de la méthode L-H pour obtenir les bornes finales. Nous expliquons comment les tracés de contours permettent de résoudre le problème de la convergence en montrant comment les contraintes peuvent être réglées de manière à être prises en compte après chaque itération. Ce procédé nous protège contre la convergence trop lente ou la non-convergence, lorsqu'on présume que les avantages attendus ne valent pas les efforts supplémentaires requis.

## 2. CONTEXTE DE L'ACES

L'ACES de 1992 a été conçue par le Bureau of the Census pour devenir un test opérationnel des méthodes d'échantillonnage, de traitement, de programmation, de saisie des données, de révision et d'estimation qui se prolongerait au-delà d'une étude pilote conduite en 1991, en prévision de la tenue de l'enquête définitive, en 1993. Les valeurs estimées des dépenses en capital correspondant

aux activités intérieures ont été publiées par groupes de branches à partir de l'enquête de 1992. En outre, les enquêtes préliminaires de 1991 et de 1992 ont fourni de précieuses données sur les dépenses en capital qui serviront aux améliorations futures des plans d'échantillonnage. L'unité de sondage de l'ACES était l'entreprise, qui pouvait comprendre plusieurs établissements. La population échantillonnée comprenait toutes les entreprises actives comptant cinq employés ou plus et oeuvrant dans tous les principaux secteurs industriels, exception faite du gouvernement. Ces secteurs comprenaient les mines, la construction, le secteur manufacturier, le transport, le commerce en gros et au détail, les finances, les services ainsi qu'une portion du secteur agricole comprenant les services agricoles, la foresterie, la pêche, la chasse et le trapage. Seules les entreprises nationales ont été incluses dans la base de sondage. Le personnel du service de recherche et de méthodologie de la Division de l'industrie du Bureau of the Census s'est chargé d'établir la base de sondage, de choisir l'échantillon et de produire les estimations.

La base d'échantillonnage de l'ACES a été construite à partir de la Standard Statistical Establishment List (SSEL) du Bureau of the Census en novembre 1992, en utilisant les données finales de 1991 pour les établissements à une seule unité (Single Unit ou SU) et les données de 1990 pour les établissements associés à des entreprises à unités multiples (multunit ou MU). Ont été exclus de la base de sondage les services gouvernementaux, les établissements de Porto Rico, de Guam, des îles Vierges et des îles Mariannes. Les données des registres secondaires d'identification des employeurs, qui correspondent à des établissements SU figurant sur la liste SSEL, mais associés à des établissements MU, les établissements associés à la production agricole et les ménages privés ont également été exclus de la base de sondage.

Le fichier fondé sur les établissements a été fondé en un fichier fondé sur les entreprises. En outre, les codes à quatre chiffres de la classification type des industries (Standard Industrial Classification ou SIC) correspondant à chaque entreprise ont été remplacés par des codes de catégorie ACES. Les 80 catégories ACES étaient constituées soit de codes SIC à trois chiffres ou de combinaisons de ces codes. La base de sondage de l'ACES comprenait environ deux millions d'entreprises.

## 3. APPLICATION DE LA METHODE L-H À L'ACES

L'univers de l'enquête a été divisé en deux strates principales. La strate I a été définie arbitrairement comme une strate à tirage complet constituée de grandes entreprises comptant plus de 500 employés et disposant d'un actif de plus de 100 millions de dollars. Les entreprises de la strate I n'ont pas été classées dans une industrie ACES. Pour les totaux estimés de la paye utilisés dans le calcul de la taille des échantillons du niveau industriel, les entreprises de la strate I pouvaient contribuer à plus d'une industrie ACES, selon le nombre des branches ACES différentes pour lesquelles elles possédaient des feuilles de paye, tel qu'indiqué dans la liste SSEL.



# Utilisation de la méthode de Lavallée et Hidiroglou pour le calcul des limites de stratification aux fins de l'enquête annuelle sur les dépenses en capital du Bureau of the Census

## JOHN G. SLANTA et THOMAS R. KRENZKE<sup>1</sup>

### RÉSUMÉ

On a utilisé la méthode de Lavallée-Hidiroglou (L-H) pour le calcul des bornes de stratification dans le cadre de l'enquête annuelle du Bureau of the Census sur les dépenses en capital (Annual Capital Expenditures Survey, ou ACES), afin de stratifier une portion de l'univers de cette enquête aux fins de l'étude pilote et des enquêtes préliminaires subséquentes. Cette méthode itérative minimise la taille de l'échantillon tout en fixant le niveau souhaité de fiabilité en établissant des bornes appropriées. Toutefois, deux problèmes se sont posés lors de son application. D'abord, des bornes de départ différentes ont engendré des bornes finales différentes. Deuxièmement, la convergence vers les bornes localement optimales a été lente; autrement dit, le nombre d'itérations était élevé et la convergence n'était pas garantie. Le présent article décrit les difficultés rencontrées avec la méthode L-H et montre comment elles ont été résolues de manière à ce que cette méthode fonctionne correctement pour l'ACES. Nous décrivons notamment comment les traces de contours ont été établies et utilisées afin d'illustrer l'importance des problèmes relevés lors de l'utilisation de la méthode L-H. Nous décrivons en outre les changements apportés à la méthode L-H afin d'en faire un moyen utile de détermination des bornes de stratification pour l'ACES.

MOTS CLÉS: Convergence; traces de contours; enquêtes économiques.

### 1. INTRODUCTION

Le plan de sondage de l'enquête annuelle du Bureau of the Census sur les dépenses en capital (Annual Capital Expenditures Survey ou ACES) a pour objectifs principaux de procurer les niveaux de fiabilité souhaités à l'aide de méthodes faisables au plan opérationnel tout en respectant les contraintes budgétaires imposées. Pour répondre à ces objectifs, nous avons mis au point un plan d'échantillonnage aléatoire simple stratifié fondé sur une version modifiée de la méthode de détermination des bornes de stratification de Lavallée et Hidiroglou (L-H) (1988). Cette méthode de stratification pour les populations asymétriques donne des bornes optimales en minimisant la taille de l'échantillon total, compte tenu d'un coefficient de variation (c.v.) souhaité. Les personnes qui sont chargées d'une enquête spécialisée et qui ont accès à une variable de stratification unique peuvent tirer parti de la facilité d'exécution de cette méthode et des économies qu'elle permet de réaliser.

Nous avons examiné plusieurs articles qui portaient sur d'autres méthodes de détermination des bornes des strates fondées sur la taille. Hess, Sethi et Balakrishnan (1966) ont comparé plusieurs méthodes de stratification. La population de Dahlenius et Hodges (Cochran 1977, p. 129) a été jugée facile à mettre en oeuvre dans notre cas, mais nous l'avons rejetée au départ puisqu'elle ne tient pas compte des strates à tirage complet au moment de sa mise au point. La méthode de Sethi (1963) utilisant les distributions normales n'a pas été retenue parce que nous avons jugé qu'il serait encombrant d'identifier la distribution et

que l'utilisation des distributions normales pour chacune des 80 types d'industrie examinés dans le cadre de l'enquête ne constituerait pas une solution optimale. La règle de l'égalité du produit des poids et de la portée des strates proposée par Eckman (1959) nous a semblé exiger des calculs plutôt rébarbatifs.

La méthode L-H est donc apparue la plus propice à nos besoins. Conçue précisément pour les populations asymétriques, une situation qu'on observe souvent avec des enquêtes économiques, cette méthode calcule une borne qui définit la strate à tirage complet, ainsi que les bornes optimales pour les strates à tirage partiel. En outre, elle crée parfois une strate à tirage complet supplémentaire si, sous l'effet de la répartition de Neyman, la taille de l'échantillon de la strate est plus grande ou égale à la taille de la strate.

La méthode L-H utilise un algorithme itératif qui débute avec le calcul ou la détermination arbitraire des bornes initiales des strates. Ensuite, on procède au calcul des paramètres tels que la taille, la moyenne et la variance des strates. Ces paramètres sont incorporés dans les formules de calcul des bornes, elles-mêmes dérivées d'une minimisation de la taille de l'échantillon en fonction d'une valeur souhaitée du c.v. Si les nouvelles bornes ne convergent pas, les paramètres des strates sont calculés à nouveau pour la strate nouvellement définie. Le cycle se poursuit jusqu'à ce que les bornes convergent.

Schneeberger (1979) a abordé le problème de la détermination des bornes de stratification optimales. Il a montré dans son article que lorsqu'on aborde ce problème sous l'angle d'un programme non linéaire, résolu par une méthode

<sup>1</sup> John G. Slanta, Manufacturing and Construction Division; Thomas R. Krenzke, Decennial Statistical Studies Division, U.S. Bureau of the Census, Washington, D.C. 20233, U.S.A.



nos modèles pour les erreurs de non-appariements erronés donnent une approximation raisonnable de la réalité, du moins dans l'application censitaire dont il est question ici. Nous avons analysé les données du système triple de St. Louis en estimant le taux d'appariement de l'EBA. Les taux d'appariement pourraient ne pas être homogènes avec différentes strates de la population. C'est pourquoi nous suggérons l'utilisation des données de l'EBA associées à la même strate d'échantillonnage. En §3, nous avons élaboré une formule pour un recensement à  $k$  échantillons et cette approche peut s'appliquer à un recensement à  $k$  échantillons où  $k \geq 4$ .

## BIBLIOGRAPHIE

- BISHOP, Y.M.M., FIENBERG, S.E., et HOLLAND, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: M.I.T. Press.
- CHEN, T.T. (1979). Log-linear models for categorical data with misclassification and double sampling. *Journal of American Statistical Association*, 74, 481-488.
- CORMACK, R.M. (1968). The statistics of capture-recapture methods. *Oceanography and Marine Biology, Annals Review*, 6, 455-506.
- DARROCH, J.N. (1958). The multiple-recapture census, I: estimation of a closed population. *Biometrika*, 45, 343-359.
- DARROCH, J.N., FIENBERG, S.E., GLONEK, G.F.V., et JUNKER, B.W. (1993). A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. *Journal of American Statistical Association*, 88, 1137-1148.
- DING, Y. (1990). Capture-recapture census with uncertain matching. Thèse de doctorat, Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania.
- DING, Y., et FIENBERG, S.E. (1994). Estimation de système dual du sous-dénombrement dans le recensement lorsqu'il y a erreur d'appariement. *Techniques d'enquête*, 20, 155-165.
- FAY, R.E., PASSEL, J.S., ROBINSON, J.G., et COWAN, C.D. (1988). The coverage of population in the 1980 census. Bureau of the Census, U.S. Department of Commerce.
- FIENBERG, S.E. (1972). The multiple recapture census for closed populations and incomplete  $2^k$  contingency tables. *Biometrika*, 59, 591-603.
- HOGAN, H., et WOLTER, K. (1988). Mesure de l'erreur dans une enquête post-censitaire. *Techniques d'enquête*, 14, 105-124.
- MULRY, M.H., DAJANI, A., et BIEMER, P. (1989). The Matching Error Study for the 1988 Dress Rehearsal. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 704-709.
- RAO, C.R. (1957). Maximum likelihood estimation for the multinomial distribution. *Sankhyā*, 18, 139-148.
- SANATHANAN, L. (1972). Estimating the size of a multinomial population. *Annals of Mathematical Statistics*, 43, 142-152.
- SEBER, G.A.F. (1982). *The Estimation of Animal Abundance and Related Parameters*. New York: MacMillan.
- ZASLAVSKY, A.M., et WOLFGANG, G.S. (1993). Triple System Modeling of Census, Post-Enumeration Survey and Administrative List Data. *Journal of Business and Economic Statistics*, 11, 279-288.

hiérarchiques, donc ne puissent se comparer directement, il semble raisonnable de retenir l'estimation la plus faible et la plus stable qui n'intègre pas l'hétérogénéité.

Darroch et coll. (1993) ont analysé quatre sous-strates de la strate 11. Les deux variables de classification croisée des quatre sous-strates O2, R2, O3 et R3 étaient: propriétaire ou locataire d'une maison, et 20 à 29 ans ou 30 à 44 ans. Les données des quatre sous-strates apparaissent au tableau 8, où 1 indique la présence et 0 l'absence dans l'échantillon. Nous avons réanalysés ces données aux fins de comparaison. Le tableau 9 et le tableau 10 donnent une estimation des deux modèles d'hétérogénéité. Comme on l'a indiqué plus haut, le taux d'appariement élevé débouche sur des estimations et un ajustement analogues pour les modèles qui incorporent l'erreur d'appariement. Le modèle à quasi-symétrie partielle donne un ajustement sensiblement plus précis que le modèle quasi-symétrique. Les meilleurs ajustements concernent les sous-strates R2 et R3. Si on additionne les estimations de  $N$  pour les quatre sous-strates, le total pour la version à quasi-symétrie partielle du modèle à erreur d'appariement donne  $N = 2980,8$ , soit 16% de plus que l'estimation obtenue avec le modèle groupé du tableau 7. Bien sûr, l'erreur-type de l'estimation augmente du même ordre de grandeur.

Tableau 8

Données à trois échantillons pour quatre sous-strates de la strate 11  
Source: tableau 2, Darroch et coll. (1993)

Échantillon				Sous-strate			
C	P	A	O2	R2	O3	R3	
0	0	1	59	43	35	43	0
0	1	0	8	34	10	24	0
0	1	1	19	11	10	13	1
1	0	0	31	41	62	32	1
1	0	1	19	12	13	7	1
1	1	0	13	69	36	69	1
1	1	1	79	58	91	72	1

Tableau 9

Estimations pour une quasi-symétrie totale				Sous-strate			
EP de Darroch et coll. (1993)		N (E.-T.)		Ajustement (d.l.)		Ajustement (d.l.)	
EP avec modèle d'appariement (2)		N (E.-T.)		Ajustement (d.l.)		Ajustement (d.l.)	
780,83 (294,81)	11,70 (2)	777,98 (293,99)	11,69 (2)				
394,34 (56,45)	41,09 (2)	391,14 (55,29)	41,02 (2)				
765,45 (254,57)	25,99 (2)	759,97 (252,44)	25,98 (2)				
361,83 (47,33)	59,31 (2)	358,71 (46,20)	59,22 (2)				
R3							

Dans le présent article, nous proposons des modèles pour les erreurs d'appariement et des modèles pour l'estimation de la population totale et du sous-dénombrement du recensement dans un recensement à échantillon multiple. Nous avons illustré les méthodes proposées en procédant à une nouvelle analyse des données sur la couverture du recensement obtenues lors de la répétition générale de 1988 à St. Louis. Nous envisageons pour cela deux sources d'information, soit les données d'une étude sur l'erreur d'appariement (BEA) et celles d'un système triple où chaque sujet a fait l'objet d'une classification croisée selon sa présence ou son absence dans un des trois échantillons suivants: recensement, enquête postcensitaire (échantillon P) et liste administrative complémentaire. Nous avons intégré la formule type du modèle log-linéaire de Fienberg (1972) à notre méthode d'estimation pour tenir compte de la dépendance statistique et des erreurs d'appariement, et afin de disposer d'un test d'adéquation formelle pour les divers modèles. Notre méthode est valable pour n'importe quel modèle de type log-linéaire et nous avons montré comment les modèles d'hétérogénéité peuvent être incorporés à cette approche afin de tenir compte à la fois des erreurs d'appariement et d'un potentiel de saisie hétérogène. Nos modèles de l'erreur d'appariement supposent des taux appariements négligeables. L'analyse de sensibilité effectuée par Ding (1990) montre que lorsque le taux de non-appariements erronés et le taux de faux appariements ont le même ordre de grandeur, le biais d'appariement résulte principalement du taux de faux non-appariements (lire aussi Fay, Passel, Robinson et Cowan 1988, p. 53). On le doit à une forte probabilité de saisie pour le recensement et l'enquête postcensitaire, si bien qu'une variation analogue des taux de non-appariements erronés et de faux appariements exerce sensiblement plus d'influence sur les non-appariements erronés que sur les faux appariements. Dans le cas des données censitaires de l'essai effectué en 1986 à Los Angeles, Ding et Fienberg (1994) estiment le taux de non-appariements erronés et le taux de faux appariements à environ 0,5% et 0,8%, respectivement. Compte tenu de ces résultats empiriques, il y a lieu de croire que

## 6. RÉCAPITULATION

Estimations pour une quasi-symétrie partielle				Sous-strate			
EP de Darroch et coll. (1993)		N (E.-T.)		Ajustement (d.l.)		Ajustement (d.l.)	
EP du modèle à erreur d'appariement (2)		N (E.-T.)		Ajustement (d.l.)		Ajustement (d.l.)	
605,66 (212,63)	7,51 (1)	601,44 (210,93)	7,52 (1)				
652,34 (205,12)	0,04 (1)	646,59 (202,58)	0,04 (1)				
1124,00 (473,26)	8,27 (1)	1126,90 (476,54)	8,22 (1)				
611,78 (200,82)	2,92 (1)	605,91 (198,26)	2,92 (1)				
R3							

Tableau 10



on observe un écart de 0,8738% dans l'estimation de  $N$  combiné à un taux de non-appariement de 0,363%. En supposant que la différence dans l'estimation de  $N$  varie de façon à peu près linéaire avec le taux de non-appariement, si le taux de non-appariement avait été de 10%, à savoir s'il y avait eu appariement à 90%, l'écart entre l'estimation habituelle du maximum de vraisemblance de  $N$  et notre propre estimation aurait été de 26%.

Tableau 6  
Données du système dual pour la strate 11 de St. Louis

Échantillon $P$	Recensement		
	Présent	Absent	Total
Présent	487	129	616
Absent	217	-	704
Total			

Le tableau 6 présente les données habituelles du système dual pour la strate 11 de St. Louis. Le nombre de gens dénombrés lors du recensement et dans l'échantillon  $P$  est  $y_{11} = 300$ , le nombre de personnes dans le recensement seulement est  $y_{12} = 217$  et le nombre de sujets dans l'échantillon  $P$  seulement est  $y_{21} = 129$ . Le dénombrement total s'élève à  $y_{1+} = y_{11} + y_{12} = 704$  pour le recensement et à  $y_{+1} = y_{11} + y_{21} = 616$  pour l'échantillon  $P$ . L'estimation du système dual est  $DSE = y_{1+}y_{+1}/y_{11} = 893$  (p. 232, Bishop, Fienberg et Holland 1975) et la variance estimée de  $DSE$  est  $Var(DSE) = y_{1+}y_{+1}y_{12}y_{21}/y_{11}^3 = 105,4$  (p. 233, Bishop et coll. 1975). L'erreur-type s'établit à  $SE(DSE) = 10,27$ .

Le sous-dénombrement pour l'estimation de la population  $DSE$  est  $(DSE - y_{1+}) / DSE \times 100\% = 21,17\%$ . Avec le modèle permettant le meilleur ajustement, le sous-dénombrement s'élève à  $(N - y_{1+}) / N = 55,97\%$  pour l'estimation  $N = 1599$ , en supposant l'absence d'erreur d'appariement, et à  $55,58\%$  pour  $N = 1585$ , selon le modèle (2) de l'erreur d'appariement. Il existe donc un biais à la hausse de  $55,97\% - 55,58\% = 0,39\%$  quand on ne tient pas compte des erreurs d'appariement. Ce résultat s'approche beaucoup des 0,37% calculés par Ding et Fienberg (1994) à partir des données du recensement expérimental de 1986 à Los Angeles au moyen d'un taux d'appariement pour deux échantillons de 99,4734%. Le taux d'appariement utilisé ici avec les données de St. Louis est de 99,6637%. Nos estimations révèlent que la population d'adultes de sexe masculin et de race noire vivant en milieu urbain, choisie pour la répétition générale de St. Louis, a été l'objet d'un sous-dénombrement important lors du recensement et que l'estimateur habituel à système dual ou à saisie et resaisie sous-estime gravement le sous-dénombrement. Un trois-système échantillon, qualitativement différent, pourrait donner de bons résultats pour ce groupe démographique. Des probabilités de saisie homogènes font partie des hypothèses de l'approche type à l'estimation d'une population fermée. Darroch et ses collaborateurs (1993) ont mis au point un modèle quasi-symétrique et un modèle à

quasi-symétrie partielle permettant de faire varier le potentiel de saisie des sujets. Le premier modèle suppose que les trois échantillons présentent le même genre d'hétérogénéité, tandis que le modèle à quasi-symétrie partielle présuppose que deux échantillons ont la même hétérogénéité mais pas le troisième. Il s'agit d'un modèle raisonnable puisque le troisième échantillon est qualitativement très différent du recensement et de l'EP. Ce modèle équivaut à une combinaison de la dépendance et de l'hétérogénéité. Pour les probabilités par cellule multinomiale,  $y$  compris la cellule manquante  $R = (r_{111}, r_{112}, \dots, r_{222})$ , il s'agit de deux modèles log-linéaires de type  $R = AB$  assortis d'une matrice de plan d'expérience  $A$  appropriée et d'un vecteur des paramètres  $\beta$ . Darroch et ses collaborateurs (1993) donnent les matrices de plan d'expérience pour les deux modèles.

Tableau 7  
Modèles à potentiel de saisie hétérogène

Modèle	log-linéaire			Quasi-symétrie		Quasi-symétrie partielle
	BP de Darroch et coll. (1993)	Ajustement (d.l.)	$N(E, T)$	Ajustement (d.l.)	BP du modèle à erreur d'appariement (2)	
	1923,63 (216,84)	133,54 (2)	1906,61 (213,47)	133,50 (2)	2576,54 (413,28)	11,70 (1) 2557,08 (409,39) 11,72 (1)

La méthode que nous proposons intègre aisément le potentiel de saisie hétérogène, ce qui permet d'estimer la taille de la population, en supposant un modèle d'hétérogénéité pour le tableau 1 et en adoptant l'estimation conditionnelle de vraisemblance (Sanathanan 1972). Le tableau 7 présente les estimations obtenues après ajustement du modèle quasi-symétrique et du modèle à quasi-symétrie partielle aux données de la strate 11. Encore une fois, les erreurs d'appariement de l'analyse n'ont pas d'effet sensible, en raison du taux d'appariement élevé. Le modèle à quasi-symétrie partielle permet un meilleur ajustement que le modèle quasi-symétrique, signe que l'hétérogénéité est plausible et que la LAC paraît avoir une hétérogénéité différente. Le fait que le modèle à indépendance totale s'ajuste mal pourrait aussi s'expliquer en partie par la dépendance entre les échantillons (en particulier entre le recensement et l'échantillon  $P$ ), et en partie par le potentiel de saisie hétérogène.

Le modèle à quasi-symétrie partielle intègre la dépendance [CP], si bien qu'il constitue une solution de rechange au modèle [CP] [PA] du tableau 5. Les deux modèles permettent un ajustement analogue des données, mais aboutissent à des estimations radicalement différentes de  $N$ , le modèle qui incorpore l'hétérogénéité produisant une estimation nettement plus élevée, assortie d'une estimation de l'erreur-type estimative beaucoup plus importante. Il se pourrait donc que les paramètres d'hétérogénéité soient très instables et, bien que les deux modèles ne soient pas



$\theta = \alpha^2$  et  $\theta = \alpha^2 = 99.3285\%$ . D'après d'autres renseignements qualitatifs, le taux d'appariement paraît anormalement élevé. En outre, le taux d'erreur de l'appariement entre le recensement et la LAC est sans doute plus élevé que le taux d'erreur de l'appariement entre le recensement et l'échantillon *P*. Faute de meilleurs renseignements quantitatifs, nous avons néanmoins décidé de les utiliser dans les calculs qui suivent.

Tableau 5  
Estimations selon divers modèles

Modèle	EP ordinaire		EP avec modèle à erreur d'appariement (2)	
	Ajustement (d.l.)	N (E.-T.)	Ajustement (d.l.)	N (E.-T.)
log-linéaire	[C] [P] [A]	1091.48 (11.24)	1083.58 (10.93)	244.56 (3)
	[CP] [A]	1204.14 (23.31)	1194.73 (22.86)	87.30 (2)
	[PA] [C]	1108.34 (13.77)	1100.03 (13.40)	244.53 (2)
	[CA] [P]	1068.87 (10.47)	1061.09 (10.10)	226.42 (2)
	[CP] [CA]	1271.11 (52.55)	1256.77 (50.97)	84.37 (1)
	[CP] [PA]	1598.88 (106.26)	1585.03 (104.93)	15.88 (1)
	[CA] [PA]	1080.47 (13.38)	1072.19 (12.88)	226.44 (1)
	[CP] [CA] [PA]	2360.82 (363.25)	2309.55 (352.36)	— (0)
	[CP] [CA] [P]	1080.47 (13.38)	1072.19 (12.88)	226.44 (1)
	[CP] [CA] [PA]	1598.88 (106.26)	1585.03 (104.93)	15.88 (1)

Le tableau 5 donne l'estimation de la taille de la population pour divers modèles log-linéaires, avec l'estimation de l'erreur-type et des statistiques d'adéquation. On a calculé l'erreur-type selon la méthode delta envisagée par Fienberg (1972). L'hypothèse d'indépendance entre le recensement et l'échantillon *P* a été contestée dans le cadre de l'ESD. La méthode du système dual ne nous permet de vérifier cette hypothèse et d'ajuster la dépendance potentielle que de façon limitée, alors que ces deux aspects peuvent être réglés au moyen de modèles log-linéaires pour trois échantillons ou plus. Quatre modèles énumérés au tableau 5 supposent l'indépendance entre le recensement et l'échantillon *P*; le modèle de totale indépendance [C] [P] [A], [PA] [C], [CA] [P], et [CA] [PA]. Aucun d'eux ne permet un bon ajustement des données. Les trois modèles avec terme d'interaction entre le recensement et l'échantillon *P*, [CP] [A], [CP] [CA] et [CP] [PA], assurent un meilleur ajustement. L'ajout d'un terme d'interaction reliant le recensement à la LAC n'améliore que légèrement l'ajustement du modèle [CP] [CA] par rapport à [CP] [A], signe que le recensement et l'échantillon *P* sont presque indépendants de la LAC. Le modèle [CP] [PA] est celui qui procure le meilleur ajustement, ce qui semble indiquer que l'hypothèse habituelle d'indépendance de l'ESD n'est pas valable et qu'il existe un lien entre l'échantillon *P* et la LAC. L'ajustement obtenu avec le modèle (2) de l'erreur d'appariement est légèrement meilleur pour les sept modèles log-linéaires insaturés, mais à peine, en raison du fort taux d'appariement des données de la répétition générale du recensement américain de 1988. Pour le modèle [CP] [PA],

On peut analyser les données d'un tel système triple au moyen du modèle (2) des erreurs d'appariement et des données d'une étude sur l'erreur d'appariement (EBP ou étude de réappariement) distincte, associées aux mêmes strates échantillonnées à posteriori. L'EBP est l'une des opérations effectuées par le Bureau of the Census pour évaluer l'EP. On applique typiquement à un échantillon de cas en recourant à des procédures plus développées et à du personnel hautement qualifié, et en procédant à de nouvelles entrevues afin d'estimer le biais associé au processus de couplage antérieur. Dans leur analyse de l'étude sur l'erreur d'appariement entreprise lors d'un recensement expérimental, en 1986, à Los Angeles, Hogan et Wolter (1988) signalent que le réappariement avait été effectué indépendamment de l'appariement original et qu'on avait établi l'écart entre les résultats de l'appariement et du réappariement. À cause de l'approche intensive utilisée, les auteurs estiment que les résultats du réappariement illustrent un couplage véritable, et que les différences entre les données des deux exercices représentent le biais qui fausse les résultats de l'appariement original. Par conséquent, les données de l'EBP permettent d'estimer le taux d'erreur lors de l'appariement initial. Mulry, Dafani et Biemer (1989) ont rédigé un rapport sur l'EBP pour la répétition générale de 1988, dans lequel ils fournissent les résultats du réappariement pour les trois sites d'essai. Le tableau 4 reproduit les données du rapport qui nous intéressent.

Tableau 4  
Etude de réappariement de St. Louis: échantillon *P*  
Source: Mulry, Dafani et Biemer (1989)

Classification à l'appariement initial				Classification au réappariement			
Apparié	Non apparié	Indéterminé	Total	Apparié	Non apparié	Indéterminé	Total
2,667	7	427	30	2,676	441	58	3,175
Apparié	8	466	27	Non apparié	9	20	Indéterminé
2,682	466	27	0	2,676	441	58	3,175
Total	2,676	441	58	Total	2,676	441	58

Soit  $\alpha$ , le taux d'appariement entre les échantillons *C* et *P* et  $\gamma = 1 - \alpha$ , le taux d'erreur du non-appariement. Nous présumons que le réappariement ne comporte aucune erreur. À partir des données du tableau 4, on peut estimer  $\alpha$  avec  $\alpha = 2667/(2667 + 9) = 99.6637\%$  et  $\gamma$  avec  $\gamma = 1 - \alpha = .3363\%$ . Le paramètre  $\theta$  représente le taux d'appariement à trois échantillons pour l'échantillon *C*, l'échantillon *P* et la LAC. Deux appariements sont nécessaires à une classification correcte (1, 1) des trois échantillons, par exemple le premier entre l'échantillon *C* et l'échantillon *P* et le second entre l'échantillon *P* et la LAC. Puisque l'appariement entre le recensement et la LAC n'a pas été évalué, nous supposons que les deux couplages sont indépendants l'un de l'autre et que le taux d'appariement entre l'échantillon *P* et la LAC est identique à celui entre l'échantillon *C* et l'échantillon *P*. On peut donc utiliser

4.2 Estimation de la taille de la population

Examinons maintenant le taux d'appariement des divers modèles. Pour estimer la population, on procède de la façon suivante. Tout d'abord, à l'instar de Sanathanan (1972), on estime le maximum de vraisemblance des paramètres du terme  $u$  au moyen de  $l_c$ , la vraisemblance conditionnelle associée au tableau 2 étant donné  $n$ ,

$$l_c = n! \prod_{\{(ijk) \neq (222)\}} \frac{(q_{ijk})^{x_{ijk}}}{x_{ijk}!},$$

où  $n = \sum_{\{(ijk) \neq (222)\}} x_{ijk} = m_{ijk}/n$ . Sanathanan (1972) montre que dans des conditions de régularité convenables, les estimations conditionnelles et inconditionnelles du maximum de vraisemblance sont à la fois cohérentes et présentent la même distribution normale asymptotique. Quand on supprime les paramètres redondants du terme  $u$  au moyen des contraintes liées au modèle log-linéaire du tableau 1, la difficulté consiste à établir le maximum de  $l_c$ , sous réserve de la seule contrainte que voici:

$$\sum_{\{(ijk) \neq (222)\}} m_{ijk} = n.$$

Sur le plan numérique, il s'agit d'un problème d'optimisation sans limite linéaire. Rao (1957) a examiné les conditions de régularité qui débouchent sur une estimation unique du maximum de vraisemblance pour les paramètres, avec une distribution multinomiale. Les conditions qu'il a établies sont satisfaites par la paramétrisation de  $\{q_{ijk}\}$ . Une fois que les estimations conditionnelles du maximum de vraisemblance des paramètres du terme  $u$  sont connues, on se sert du modèle log-linéaire établi pour le tableau 1 afin d'obtenir les estimations conditionnelles du maximum de vraisemblance pour  $\{l_{ijk}\}$ , soit le chiffre prévu par la cellule du tableau 1,  $y$  compris le chiffre prévu pour la cellule manquante. L'estimation de  $N$  est donc

$$N = n + m_{222},$$

S'il n'y a pas d'erreur d'appariement, puisque  $\alpha_1 = \alpha_2 = 1$  au modèle (3),  $m_{ijk} = l_{ijk}$ . Par conséquent,

$$N = \sum_{\{(ijk)\}} l_{ijk}.$$

ce qui revient à la méthode d'estimation du problème classique du recensement à resaisie multiple énoncé par Fienberg (1972) avec les modèles log-linéaires proposés par cet auteur. Comme on l'a indiqué précédemment, on a spécifié un modèle log-linéaire pour le tableau 1 et on considère que les observations s'inscrivent au tableau 2, dont le modèle paramétrique pour le chiffre prévu par cellule est défini par le modèle log-linéaire et par le modèle retenu pour les erreurs d'appariement. Pour savoir si un modèle log-linéaire spécifié pour le tableau 1 convient, on peut appliquer les

tests habituels d'adéquation de Pearson et du ratio de vraisemblance,  $X^2$  et  $G^2$ , décrits dans Fienberg (1972), au tableau 2. Chaque statistique se caractérise par une distribution  $\chi^2$  asymptotique selon l'hypothèse nulle que le modèle convient avec  $2^k - 1$  — (nombre de paramètres indépendants du modèle) degrés de liberté.

5. ANALYSE DES DONNÉES DE LA RÉPÉTITION GÉNÉRALE DU RECENSEMENT DE 1988 ST. LOUIS

Le U.S. Bureau of the Census se sert de l'estimation à système dual (ESD) articulée sur le recensement type à deux échantillons pour évaluer la couverture du recensement depuis 1950. En 1988, le Bureau a procédé à une répétition générale en prévision du recensement décennal de 1990 à trois endroits: St. Louis (Missouri), Columbia (Missouri) et l'ouest de l'Etat de Washington. Zaslavsky et Wolfgang (1993) présentent les données d'un sous-groupe de la population extrait de l'enquête postcensitaire (EPC) de la répétition générale de St. Louis. Ce sous-groupe est essentiellement composé d'adultes de race noire et de sexe masculin vivant en milieu urbain, dont on croit la population sous-estimée par les méthodes à système dual. Les données résultantes viennent de trois sources: l'échantillon C est celui du recensement proprement dit; l'échantillon P dérive de l'EPC et on a recouru à une troisième source d'information, en l'occurrence la liste administrative complémentaire (LAC) tirée des dossiers administratifs de divers services d'Etat et services fédéraux (y compris sécurité de l'emploi, permis de conduire, impôt, service sélectif et administration des anciens combattants) avant le recensement. L'échantillon C et l'échantillon P procurent les données nécessaires à l'ESD habituelle ou à l'approche par saisie et resaisie. On peut combiner les données de la LAC à celles du recensement et de l'échantillon P pour effectuer une analyse sous l'angle d'un triple échantillon, même si le but original de l'exercice était d'améliorer la couverture de l'échantillon P. Le tableau 3 présente les données à trois échantillons relatives à la strate d'échantillonnage 11 de l'EPC pour St. Louis, obtenues par groupement des données originales du tableau 1 de Zaslavsky et Wolfgang (1993) pour quatre strates à posteriori définies par les paramètres propriétaires/locataire  $\times$  20-29, 30-44 ans.

Tableau 3

Données à trois échantillons pour la strate 11 de St. Louis

LAC	Recensement			
	Présent		Absent	
	échantillon P		échantillon P	
	Présent	Absent	Présent	Absent
Present	300	51	166	76
Absent	187	180	—	—



$$F = \bigcup_{\{i,j,k\}} (E_{ijk} \cap F).$$

Selon le modèle (3), il n'existe que quatre transitions en vertu desquelles  $F$  se réalise, en l'occurrence:

$$\begin{aligned} (1,1,1) &\mapsto \{(1,0,0),(0,1,1)\}, \\ (1,1,0) &\mapsto \{(1,0,0),(0,1,0)\}, \\ (1,0,1) &\mapsto \{(1,0,0),(0,0,1)\}, \\ (1,0,0) &\mapsto \{(1,0,0)\}. \end{aligned}$$

Par conséquent,

$$F =$$

$$(E_{111} \cap F) \cup (E_{112} \cap F) \cup (E_{121} \cap F) \cup (E_{122} \cap F).$$

Etant donné la définition de la probabilité par cellule des deux tableaux,  $p(F) = p_{122}$  et  $p(E_{ijk}) = r_{ijk}$ . En vertu des hypothèses du modèle (3),  $p(F | E_{111}) = (1 - \alpha_1)/3$ ,  $p(F | E_{112}) = p(F | E_{121}) = \alpha_2$  et  $p(F | E_{122}) = 1$ . Puisque  $E_{111} \cap F$ ,  $E_{112} \cap F$ ,  $E_{121} \cap F$  et  $E_{122} \cap F$  sont quatre possibilités qui s'excluent mutuellement pour que  $F$  se réalise,

$$p_{122} = p(E_{111} \cap F) + p(E_{112} \cap F)$$

$$+ p(E_{121} \cap F) + p(E_{122} \cap F)$$

$$= p(F | E_{111}) \cdot p(E_{111}) + p(F | E_{112}) \cdot p(E_{112})$$

$$+ p(F | E_{121}) \cdot p(E_{121}) + p(F | E_{122}) \cdot p(E_{122})$$

$$= \frac{1 - \alpha_1}{3} r_{111} + (1 - \alpha_2) r_{112} + (1 - \alpha_2) r_{121} + r_{122}.$$

On peut dériver les autres probabilités par cellule du tableau 2 de la même manière, si bien qu'on obtient

$$p_{111} = \alpha_1 r_{111},$$

$$p_{112} = \frac{1 - \alpha_1}{3} r_{111} + \alpha_2 r_{112},$$

$$p_{121} = \frac{1 - \alpha_1}{3} r_{111} + \alpha_2 r_{121},$$

$$p_{211} = \frac{1 - \alpha_1}{3} r_{111} + \alpha_2 r_{211},$$

$$p_{122} = \frac{1 - \alpha_1}{3} r_{111} + (1 - \alpha_2) r_{112} + (1 - \alpha_2) r_{121} + r_{122},$$

## 4. ESTIMATION DE LA TAILLE DE LA POPULATION

### 4.1 Formulation du modèle log-linéaire

Aux fins d'exposition, nous nous limiterons au cas du recensement à trois échantillons, bien que l'extension à  $k$  échantillons fait précédemment, soient  $l_{ijk}$  et  $m_{ijk}$ , le chiffre par cellule prévu pour les tableaux 1 et 2, respectivement. La relation entre la probabilité par cellule et le chiffre prévu par cellule est  $l_{ijk} = r_{ijk}N$  et  $m_{ijk} = p_{ijk}N$ . Soient

$$\vec{m} = (m_{111}, m_{112}, m_{121}, m_{122}, m_{211}, m_{212}, m_{221})^T,$$

$$\vec{l} = (l_{111}, l_{112}, l_{121}, l_{211}, l_{122}, l_{212}, l_{221})^T.$$

Puisque, pour chacun des modèles proposés à la dernière partie, il existe une matrice  $M$  dont les inscriptions dépendent des paramètres de la probabilité d'appartenance pour le modèle retenu, ce qui  $\vec{p} = M \times \vec{r}$ , multiplié par  $N$  donne

$$\vec{m} = M \times \vec{l}. \quad (4)$$

La paramétrisation de  $m_{ijk}$  est simple avec tous les modèles log-linéaires indiqués pour le tableau 1. Ainsi, pour chacun des modèles suggérés par Fienberg (1972), on peut exprimer le chiffre prévu au moyen des fonctions des paramètres du terme  $u$ :

$$l_{ijk} =$$

$$g_{ijk}(u, u_1(i), u_2(j), u_3(k), u_{12}(ij), u_{13}(ik), u_{23}(jk)), \quad (5)$$

On peut ensuite dériver les paramètres de  $\{m_{ijk}\}$  à partir de (4).



Le modèle paramétrique qui exprime  $\{p_{ijk}\}$  en termes de  $\{r_{ijk}\}$  pour le cas à trois échantillons est le suivant:

$$\begin{aligned} p_{111} &= \theta r_{111}, \\ p_{112} &= \frac{3}{1-\theta} r_{111} + \theta r_{112}, \\ p_{121} &= \frac{3}{1-\theta} r_{111} + \theta r_{121}, \\ p_{211} &= \frac{3}{1-\theta} r_{111} + \theta r_{211}, \end{aligned}$$

$$\begin{aligned} p_{122} &= \frac{3}{1-\theta} r_{111} + (1-\theta)r_{112} + (1-\theta)r_{121} + r_{122}, \\ p_{212} &= \frac{3}{1-\theta} r_{111} + (1-\theta)r_{112} + (1-\theta)r_{211} + r_{212}, \\ p_{221} &= \frac{3}{1-\theta} r_{111} + (1-\theta)r_{211} + (1-\theta)r_{121} + r_{221}. \end{aligned}$$

Par conséquent

$$\vec{p} = M_2 \times \vec{r}, \quad (2)$$

où  $M_2$  est une matrice  $7 \times 7$  définie par les sept équations qui précèdent, dérivées du modèle (2). Encore une fois, la probabilité de saisie reste la même, c.-à-d.  $p_{1++} = r_{1++} = p_1$ ,  $p_{+1+} = r_{+1+} = p_2$ ,  $p_{++1} = r_{++1} = p_3$ . Pour le problème à  $k$  échantillons, soit  $p_T$ , la probabilité d'être saisi dans chaque échantillon, à savoir  $p_T = p_{111\dots 1}$ , et soit  $p_{T,\bar{z}}(h_1, h_2)$ , la probabilité par cellule correspondant à une absence dans le  $h_1$ -ième et le  $h_2$ -ième échantillons, et d'une présence dans les autres échantillons, etc. En vertu du modèle (2),  $p_T = \theta r_T$ . Lorsque  $i \leq k-2$ , la probabilité de ne pas apparaître dans les  $h_1$ -ième,  $h_2$ -ième, ..., et  $h_i$ -ième échantillons et saisis dans les autres échantillons est égale à

$$p_{T,\bar{z}}(h_1, h_2, \dots, h_i) = \theta r_{T,\bar{z}}(h_1, h_2, \dots, h_i) +$$

$$\frac{1-\theta}{i} \sum_{j=1}^k \frac{k-i+1}{r_{T,\bar{z}}(\{h_1, h_2, \dots, h_i\} \setminus h_j)}.$$

Quand  $i = k-1$ , le sujet ne fait partie que d'un échantillon. Par exemple, la probabilité de n'être saisi que dans le premier échantillon est égale à

$$p_{1,\bar{z}} = r_{1,\bar{z}} + (1-\theta) \sum_{h \neq 1} r_{1,1(h),\bar{z}} +$$

$$\frac{3}{(1-\theta)} \sum_{h_1, h_2 \geq 2} r_{1,1(h_1, h_2),\bar{z}} +$$

$$\sum_{k=1}^j \sum_{h_1, h_2, \dots, h_j \geq 2} \frac{(j+1)}{(1-\theta)} r_{1,1(h_1, h_2, \dots, h_j),\bar{z}},$$

où  $r_{1,1(h_1, h_2, \dots, h_j),\bar{z}}$  représente la probabilité par cellule du tableau original, correspondant à une présence dans le premier  $h_1$ -ième,  $h_2$ -ième, ...,  $h_j$ -ième échantillon et à une absence dans les autres échantillons. Pour des raisons de symétrie, on peut écrire l'expression pour  $p_{1(h),\bar{z}}$ , c.-à-d. la probabilité d'être observé dans le  $h$ -ième échantillon uniquement et pas dans les autres.

Il est possible d'améliorer le modèle (2) en supposant des taux d'appariement inégaux. Examinons, par exemple, deux décompositions:  $(1, 1, 1) \mapsto \{(1, 1, 0), (0, 0, 1)\}$  et  $(1, 1, 0) \mapsto \{(0, 1, 0), (1, 0, 0)\}$ . Dans les deux cas, le point commun est qu'un appariement présumé ne s'est pas réalisé. La différence réside dans le fait que, dans le premier cas, il y a couplage de deux sources d'information, comparativement à une seule dans le second. On peut raisonnablement supposer que la probabilité d'une erreur d'appariement diffère dans les deux cas, contrairement à ce que suppose le modèle (2). Ceci nous amène donc au:

**Modèle (3).** En plus des hypothèses (i) et (iii) du modèle (2), on suppose

$$\left. \begin{aligned} (1, 1, 1) &\mapsto \{(1, 1, 0), (0, 0, 1)\} \text{ avec une probabilité } \alpha_1 \\ (1, 1, 1) &\mapsto \{(0, 1, 1), (1, 0, 0)\} \text{ avec une probabilité } (1-\alpha_1)/3 \\ (1, 1, 0) &\mapsto \{(0, 1, 1), (1, 0, 0)\} \text{ avec une probabilité } (1-\alpha_1)/3 \\ (1, 1, 0) &\mapsto \{(1, 0, 1), (0, 1, 0)\} \text{ avec une probabilité } (1-\alpha_1)/3 \end{aligned} \right\}$$

$$\left. \begin{aligned} (1, 1, 0) &\mapsto \{(0, 1, 1), (1, 0, 0)\} \text{ avec une probabilité } \alpha_2 \\ (1, 1, 0) &\mapsto \{(1, 0, 1), (0, 0, 1)\} \text{ avec une probabilité } 1-\alpha_2 \\ (1, 0, 1) &\mapsto \{(1, 0, 0), (0, 0, 1)\} \text{ avec une probabilité } \alpha_2 \\ (1, 0, 1) &\mapsto \{(0, 1, 1), (0, 0, 1)\} \text{ avec une probabilité } 1-\alpha_2 \end{aligned} \right\}$$

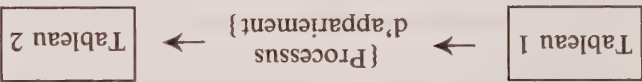
$$\left. \begin{aligned} (0, 1, 1) &\mapsto \{(0, 1, 1), (0, 0, 1)\} \text{ avec une probabilité } \alpha_2 \\ (0, 1, 1) &\mapsto \{(0, 1, 1), (0, 0, 1)\} \text{ avec une probabilité } 1-\alpha_2 \end{aligned} \right\}$$

et  $(1, 0, 0)$ ,  $(0, 1, 0)$  et  $(0, 0, 1)$  ne varient pas avec une probabilité égale à un. Selon ce modèle, on peut exprimer la probabilité par cellule  $\{p_{ijk}\}$  du tableau 2 en fonction de  $\alpha_1$ ,  $\alpha_2$  et de la probabilité par cellule du tableau 1,  $\{r_{ijk}\}$ . Pour cela, il faut tenir compte de toutes les transitions possibles qui donneront un sujet dans la cellule  $(i, j, k)$  du tableau 2. Par exemple, prenons le sujet observé  $(1, 0, 0)$ . Ce sujet se retrouve dans la cellule  $(1, 2, 2)$  du tableau 2. Soit  $F$ , l'observation d'un sujet de statut  $(1, 0, 0)$  et soit  $E_{ijk}$ , l'inclusion d'un sujet dans la cellule  $(i, j, k)$  du tableau 1. Alors,

erreur au sujet (0,1,0) donnera une seule observation telle décomposition ou combinaison une *transition*. Les transitions ne peuvent aller que du niveau 3 ou du niveau 2 au même niveau (s'il n'y a pas d'erreur d'appariement) ou à un niveau inférieur, s'il n'y a pas de faux appariements. Plus précisément, un sujet (1,1,1) peut faire la transition vers l'un des 5 jeux possibles de sujets que voici :

$$\{(1,1,1)\}, \{(1,0,0), (0,1,1)\}, \{(0,1,0), (1,0,1)\} \\ \{(0,0,1), (1,1,0)\}, \{(1,0,0), (0,1,0), (0,0,1)\}.$$

Au niveau 2, un sujet (1,1,0) pourra se décomposer en  $\{(1,0,0), (0,1,0)\}$  ou demeurer  $\{(1,1,0)\}$ . Il en ira autant des sujets  $\{(0,1,1)\}$  et  $\{(1,0,1)\}$ . De ce qui précède, on peut résumer l'effet d'une erreur d'appariement au moyen du diagramme suivant :



où le tableau 1 représente le tableau 2<sup>k</sup> incomplet original, sans erreur d'appariement, et le tableau 2 correspond au tableau 2<sup>k</sup> incomplet observé, avec erreur d'appariement. Par conséquent, la probabilité par cellule et le chiffre prévu d'une cellule du tableau 1 seront notés par  $\{r_{ijk}\}$  et  $\{l_{ijk}\}$ , et par  $\{p_{ijk}\}$  et  $\{m_{ijk}\}$  pour le tableau 2, où  $1 \leq i, j, k \leq 2$ .

### 3. QUELQUES MODÈLES POUR LES ERREURS D'APPARIEMENT

Nous présenterons maintenant des modèles qui décrivent les erreurs d'appariement. Chacun d'eux nous permet de formuler la répartition de la probabilité et du chiffre prévu des cellules associées au tableau 1.

**Modèle (1).** Outre les hypothèses d'homogénéité et de finitude expliquées en §1, nous supposons : (i) que le couplage n'entraîne pas de faux appariements; (ii) que la probabilité qu'un sujet garde son état original est égale à  $\theta$  et que la probabilité d'effectuer une transition vers un ensemble possible de sujets est égale à  $(1 - \theta)/(m - 1)$ , où  $m$  représente le nombre de jeux possibles de sujets pour lesquels la transition est autorisée. Pour le sujet (1,1,1) examiné à la dernière partie, par exemple,  $m = 5$ .

Un tel modèle permet d'exprimer les probabilités du tableau avec erreur d'appariement,  $\{p_{ijk}\}$ , d'après les probabilités du tableau sans erreur d'appariement,  $\{r_{ijk}\}$ , pour le recensement à trois échantillons de la façon suivante :

$$p_{111} = \theta r_{111}, \\ p_{112} = \frac{4}{1 - \theta} r_{111} + \theta r_{112}, \\ p_{121} = \frac{4}{1 - \theta} r_{111} + \theta r_{121},$$

$$p_{211} = \frac{4}{1 - \theta} r_{111} + \theta r_{211}, \\ p_{122} = \frac{2}{1 - \theta} r_{111} + (1 - \theta)r_{112} + (1 - \theta)r_{121} + r_{122}, \\ p_{212} = \frac{2}{1 - \theta} r_{111} + (1 - \theta)r_{112} + (1 - \theta)r_{211} + r_{212}, \\ p_{221} = \frac{2}{1 - \theta} r_{111} + (1 - \theta)r_{211} + (1 - \theta)r_{121} + r_{221}.$$

Soit

$$\vec{p} = (p_{111}, p_{112}, p_{121}, p_{122}, p_{211}, p_{212}, p_{221})^T,$$

et

$$\vec{r} = (r_{111}, r_{112}, r_{121}, r_{122}, r_{211}, r_{212}, r_{221})^T,$$

donc

$$\vec{p} = M_1 \times \vec{r}. \quad (1)$$

Dans ce cas,  $M_1$  représente une matrice  $7 \times 7$  définie par les sept équations qui précèdent, issues du modèle (1). Il est assez facile de vérifier que la probabilité de saisir un sujet dans chaque échantillon est fixe, à savoir  $p_{1++} = r_{1++} = p_1, p_{+1+} = r_{+1+} = p_2, p_{++1} = r_{++1} = p_3$ . Il doit en être ainsi puisque la probabilité de saisie dans l'échantillon ne dépend pas de la façon dont fonctionne la méthode d'appariement.

On peut aisément généraliser la formule pour  $k$  échantillons. Les opérations algébriques deviennent néanmoins fort fastidieuses quand  $k$  est élevé. Il est possible de simplifier le modèle en exigeant que les transitions descendent d'au moins un niveau, ce qui nous amène au

**Modèle (2).** En plus des hypothèses d'homogénéité et de finitude données en §1, on suppose : (i) que le couplage n'entraîne pas de faux appariements; (ii) qu'une transition ne peut se faire que vers le bas, d'au moins un niveau;

(iii) que la probabilité qu'un sujet garde son état original est  $\theta$  et que la probabilité d'une transition vers un ensemble possible de sujets est  $(1 - \theta)/(m' - 1)$ , où  $m'$  représente le nombre de combinaisons de sujets vers lesquels la transition est possible et autorisée.

Examinons d'abord le cas à trois échantillons. Un sujet (1,1,1) peut se décomposer en trois individus, soit  $(1,1,1) \mapsto \{(1,0,0), (0,1,0), (0,0,1)\}$  (le signe «  $\mapsto$  » indique qu'il y a décomposition), si les trois appariements présumés ne se concrétisent pas. L'hypothèse (ii) du modèle (2) établit que cette triple erreur a une probabilité négligeable de se reproduire, comparativement à la probabilité d'une transition où un seul appariement n'est pas effectué, si bien que  $(1,1,1) \mapsto \{(1,1,0), (0,0,1)\}$ , ou  $(1,1,1) \mapsto \{(1,0,1), (0,1,0)\}$ , ou  $(1,1,1) \mapsto \{(1,1,1), (0,0,0)\}$ .



où 3\* et 4\* représentent les sujets 3 et 4, mais associés à des renseignements incorrects, ce qui entraîne deux non-appariements erronés lors du couplage des échantillons. Partant de l'hypothèse que tous les appariements sont exacts, on obtient le tableau 2<sup>3</sup> incomplet que voici:

Tableau 2

Observations avec erreurs d'appariement					
$s_1$			$s_2$		
Présent		Absent	Présent		Absent
$s_3$			$s_3$		
Présent	0	1	Présent	0	3
Absent	0	3	Absent	0	-

Lorsqu'on compare les deux tableaux, il est facile de voir l'effet des erreurs d'appariement:

- (i) Le nombre d'observations peut augmenter dans certaines cellules et diminuer dans d'autres, si bien que les totaux marginaux, en particulier le nombre total de sujets différents observés dans les trois échantillons, peuvent varier, sous réserve de la contrainte que le nombre total d'observations dans les échantillons,  $x_{1++}$ ,  $x_{+1+}$ , et  $x_{++1}$ , reste le même. Une modification du nombre total de sujets différents dans les échantillons soulèverait un problème distinct du problème habituel posé par une erreur de classification dans l'analyse des données catégoriques, où on examine la possibilité qu'une erreur se glisse dans la répartition des sujets entre les différentes catégories (p. ex., lire Chen 1979).
- (ii) Parallèlement, la probabilité dans certaines cellules peut changer pourvu que la probabilité  $p_{1++}$ ,  $p_{+1+}$ , et  $p_{++1}$ , d'être saisie dans un échantillon demeure la même.

En raison de la complexité des erreurs d'appariement du cas à trois échantillons, nous avons besoin d'une terminologie particulière pour faciliter la description. Par rapport à l'échantillon  $s_1$ , nous dirons donc qu'un sujet se trouve à l'état 1 si on l'observe dans l'échantillon et à l'état 0 dans le cas contraire. Le triplet  $(i, j, k)$ ,  $0 \leq i, j, k \leq 1$  indiquera un sujet à l'état  $i, j$  et  $k$ , pour les échantillons  $s_1, s_2$  et  $s_3$ , respectivement. Ainsi,  $(1, 1, 0)$  dénotera un sujet qu'on observe uniquement dans  $s_1$  et  $(1, 1, 1)$ , un sujet saisi dans les trois échantillons. Le niveau d'un sujet  $(i, j, k)$  sera  $i + j + k$ , soit le nombre d'échantillons dans lequel se retrouve le sujet. Il existe donc quatre niveaux différents, 0, 1, 2 et 3. Un sujet n'aura le niveau 0 que s'il n'apparaît dans aucun échantillon, alors qu'il sera de niveau 3 quand il est présent dans les trois échantillons. Si le couplage est incorrect selon la règle d'appariement, le sujet  $(1, 1, 0)$  se décomposera en «deux sujets différents»  $(1, 0, 0)$  et  $(0, 1, 0)$ , en supposant qu'il n'y ait pas de faux appariement. Par ailleurs, le sujet  $(1, 0, 0)$  apparié par

à l'absence des  $k$  échantillons n'existe pas. L'objectif consiste à estimer le nombre de sujets de la population qui ne sont pas observés, donc qui correspondent au chiffre de la cellule manquante dans le tableau de contingence  $2^k$ . À la partie 2, nous examinerons les effets des erreurs d'appariement sur le tableau  $2^k$  incomplet observé. La partie 3 propose quelques modèles pour les erreurs d'appariement en mesure de caractériser une méthode d'appariement sujette aux erreurs. À partir de ces modèles et des hypothèses (3) et (4), nous élaborerons une procédure pour estimer la taille de la population à partir d'un modèle log-linéaire. Enfin, à la partie 5, nous nous servirons des méthodes envisagées pour analyser les données de la répétition générale de 1988 du recensement entreprise par le U.S. Bureau of the Census.

## 2. LES ERREURS D'APPARIEMENT DANS LES RECENSEMENTS À ÉCHANTILLON MULTIPLE

Nous commencerons par répartir les erreurs d'appariement en deux grandes catégories: les faux appariements et les non-appariements erronés. Pour bien comprendre la nature de ces erreurs dans un recensement à trois échantillons, prenons le cas d'un recensement à trois échantillons. Supposons qu'aucune donnée ne manque ou qu'il n'y ait aucune erreur dans la saisie de l'information sur un sujet de la population et qu'on prélève trois échantillons de la population,  $s_1, s_2$  et  $s_3$ . Supposons, par exemple, qu'on retrouve les sujets 1, 3, 4 et 7 dans l'échantillon  $s_1$ , les sujets 3, 4 et 8 dans l'échantillon  $s_2$  et les sujets 4, 9 et 10 dans l'échantillon  $s_3$ . Sous sa forme vectorielle, cette situation s'exprime comme suit:  $s_1 = (1, 3, 4, 7), s_2 = (3, 4, 8)$  et  $s_3 = (4, 9, 10)$ . Il n'y a pas d'erreurs d'appariement, pourvu que les renseignements soient complets et exacts. On obtient le tableau incomplet 2<sup>3</sup> qui suit au moyen des trois échantillons:

Tableau 1

Tableau original sans erreur d'appariement					
s <sub>1</sub>			s <sub>2</sub>		
Présent		Absent	Présent		Absent
s <sub>3</sub>			s <sub>3</sub>		
Présent	1	0	Présent	0	2
Absent	1	2	Absent	1	-

Supposons par ailleurs qu'en raison de données manquantes ou de renseignements incorrects, on observe en réalité ce qui suit:

$$s_1 = (1, 3, 4, 7), \quad s_2 = (3^*, 4^*, 8), \quad s_3 = (4, 9, 10),$$

# Estimation de la population par échantillonnage multiple et sous-dénombrement lors du recensement en présence d'erreurs d'appariement

YE DING et STEPHEN E. FIENBERG<sup>1</sup>

## RÉSUMÉ

Les auteurs examinent le recensement par saisie-resaisie multiples en assouplissant l'hypothèse classique d'un appariement parfait. Ils proposent des modèles avec erreur d'appariement permettant de caractériser les méthodes d'appariement sujettes à des erreurs. Les données observées prennent la forme d'un tableau de contingence  $2^k$  auquel manque une cellule et suivent une distribution multinomiale. Les auteurs proposent une méthode pour estimer la population. Cette approche s'applique à la fois aux modèles log-linéaires habituels pour les tableaux de contingence et aux modèles log-linéaires de l'hétérogénéité du potentiel de saisie. Enfin, les auteurs illustrent leur méthode et procèdent à une estimation en recourant à une répétition générale du recensement de 1990, effectuée en 1988 par le U.S. Bureau of the Census.

MOTS CLÉS: Recensement par saisie-resaisie; estimations de la population totale; modèles log-linéaires; erreurs d'appariement; recensement par resaisie multiple.

## 1. INTRODUCTION

La technique du recensement par resaisie multiple a été appliquée dans de nombreux domaines pour estimer l'importance d'une population fermée. Ainsi, Cormack (1968) et Seber (1982) offrent une excellente revue de nombreuses techniques utilisées. Nous envisagerons ici une séquence d'échantillons,  $s_1, \dots, s_k$ , pour laquelle les sujets du  $i$ -ième échantillon sont identifiés de façon unique, par exemple par étiquetage ou marquage, avant d'être remis dans la population (Darroch 1958). Les méthodes de recensement par resaisie multiple habituelles reposent sur les hypothèses que voici:

- (1) **Appariement idéal.** Les sujets d'une liste (source d'information, échantillon) peuvent être couplés à ceux d'une autre liste sans erreur. En d'autres termes, aucune erreur de classification ne survient lorsqu'on s'efforce de déterminer si un sujet donné a été enregistré par les deux sources d'information ou par une seule d'entre elles.
- (2) **Indépendance.** Les listes sont indépendantes l'une de l'autre, c'est-à-dire que la probabilité qu'un sujet fasse partie d'une liste ne dépend pas de l'inclusion du sujet aux listes antérieures.
- (3) **Homogénéité (potentiel de saisie identique).** Tous les sujets de la population à l'étude ont la même probabilité d'être observés (saisis) dans une liste (échantillon) quelconque.
- (4) **Finitude.** La population est «fermée». Elle ne subit aucun changement consécutivement aux naissances, aux décès, à l'émigration ou à l'immigration au cours de la période durant laquelle s'effectue l'échantillonnage.

Darroch (1958) a examiné le recensement par resaisie multiple sous l'angle de ces quatre hypothèses. De son côté, Fienberg (1972) s'est servi d'un modèle log-linéaire pour autoriser une dépendance statistique entre des types spécifiques de l'échantillon. Il a donc abandonné l'hypothèse d'indépendance. Darroch, Fienberg, Glonek et Junker (1993) ont pour leur part mis au point un modèle log-linéaire élargi qui autorise l'hétérogénéité au niveau individuel et la dépendance, mais exige au moins trois échantillons, à savoir  $k = 3$ . Plusieurs auteurs se sont penchés sur les problèmes de couplage attribuables aux erreurs d'appariement et de non-appariement dans le contexte du recensement à deux échantillons utilisé par le U.S. Bureau of the Census pour évaluer la couverture du recensement. Ding et Fienberg (1994), par exemple, ont envisagé la modélisation des erreurs d'appariement dans le recensement à deux échantillons et mis au point une méthode systématique pour estimer la population totale. L'inclusion d'un troisième échantillon (tiré des dossiers administratifs) au modèle et l'estimation de la couverture du recensement avaient déjà été envisagées par le Bureau of the Census dans le passé. On n'a pas écarté cette possibilité pour étendre et évaluer l'approche du système dual. Dans le présent article, nous examinerons les modèles d'erreur d'appariement relatifs au problème du recensement à échantillon multiple tolérant les principes de dépendance et d'hétérogénéité.

Nous supposons ici que les observations tirées des données d'un recensement par saisie multiple font partie d'une classification croisée  $2^k$ , avec ou sans le  $i$ -ième échantillon définissant la catégorie de la  $i$ -ième dimension. En vertu d'une telle classification, la cellule qui correspond

<sup>1</sup> Ye Ding, Research Scientist, Bureau of Biometrics, New York State Health Department, Concourse, Room C-144, Empire State Plaza, Albany, New York 12237, U.S.A.; Stephen E. Fienberg, Maurice Falk Professor of Statistics and Social Science, Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, U.S.A.





poids des diverses méthodes de rechange. Ce résultat laisse conclure que le choix des variables auxiliaires est important, et qu'il l'est probablement plus que le choix de la méthode de pondération. Même si les méthodes plus systématiques utilisées dans la présente recherche pour le choix des variables auxiliaires n'ont pas conduit à des améliorations importantes par rapport aux méthodes courantes des enquêtes SIPP, un choix des variables auxiliaires fondé sur l'analyse risque d'être plus productif dans d'autres études.

Lorsqu'on dispose d'un nombre important de variables auxiliaires liées à la propension à répondre, il paraît sage d'en utiliser le plus grand nombre possible aux fins de la correction de la non-réponse, en guise de sauvegarde pour la correction du biais de non-réponse. Cette stratégie générale devrait toutefois être tempérée par une évaluation soigneuse de la variation des poids obtenus, afin d'éviter une trop grande perte de précision dans les estimations de l'enquête. En outre, il conviendrait de tenir compte de la facilité de mise en vigueur des méthodes de pondération. Si, comme c'est le cas dans la présente étude, les méthodes de pondération de rechange donnent des poids et des estimations très semblables, on jugera peut-être préférable de choisir la plus simple.

REMERCIEMENTS

Nous remercions les membres du comité de lecture pour les commentaires utiles qu'ils ont formulés sur la version antérieure du présent article. Cet article porte sur des recherches effectuées pour le U.S. Bureau of the Census. Les opinions exprimées sont celles des auteurs et ne reflètent pas nécessairement celles du Bureau of the Census.

BIBLIOGRAPHIE

CHAPMAN, D.W., BAILEY, L., et KASPRZYK, D. (1986). Méthodes de compensation de la non-réponse au U.S. Bureau of the Census. *Techniques d'enquête*, 12, 161-180.

DEMING, W.E., et STEPHAN, F.F. (1942). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11, 427-444.

DEVILLE, J.-C., et SÄRNDAAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

DEVILLE, J.-C., SÄRNDAAL, C.-E., et SAUTORY, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.

JABINE, T.B., KING, K.E., et PETRONI, R.J. (1990). *Survey of Income and Program Participation (SIPP): Quality Profile*. Washington, DC: U.S. Bureau of the Census.

KALTON, G. (1983). *Compensating for Missing Survey Data*. Ann Arbor, MI: Survey Research Center, University of Michigan.

KALTON, G., et KASPRZYK, D. (1986). Le traitement des données d'enquête manquantes. *Techniques d'enquête*, 12, 1-16.

KALTON, G., LEPKOWSKI, J.M., MONTANARI, G.E., et MALIGALIC, D. (1990). Characteristics of second wave nonrespondents in a panel survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 462-467.

KALTON, G., LEPKOWSKI, J., et LIN, T. (1985). Compensating for wave nonresponse in the 1979 ISDP Research Panel. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 372-377.

KASS, G.V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29, 119-127.

KISH, L. (1992). Weighting for unequal  $P_i$ . *Journal of Official Statistics*, 8, 183-200.

LEPKOWSKI, J., KALTON, G., et KASPRZYK, D. (1989). Weighting adjustments for partial nonresponse in the 1984 SIPP Panel. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 296-301.

LITTLE, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *Revue Internationale de Statistique*, 54, 2, 137-139.

NELSON, D., McMILLLEN, D., et KASPRZYK, D. (1985). *An Overview of the SIPP, Update 1*. SIPP Working Paper No. 8401. Washington, DC: U.S. Bureau of the Census.

OH, H.T., et SCHEURER, F. (1983). Weighting adjustments for unit nonresponse. Dans *Incomplete Data in Sample Surveys, Volume 2: Theory and Bibliographies* (Eds. W.G. Madow, I. Olkin, et D. Rubin), 143-184. New York: Academic Press.

PENNELL, S.G., et LEPKOWSKI, J.M. (1992). Panel conditioning effects in the Survey of Income and Program Participation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 556-571.

RIZZO, L., KALTON, G., et BRICK, M. (1994). Weighting adjustments for panel nonresponse in the SIPP. Rapport final présenté au U.S. Bureau of the Census pour le SIPP Panel Nonresponse Project.



nous paraît vraisemblable que des facteurs autres que le biais de non-réponse du panel soient largement responsables de l'ampleur de ces différences normalisées fondées sur ces sources de données principalement administratives mettent peut-être en lumière certains problèmes importants liés à la qualité des données (des enquêtes SIP, des sources de données de référence ou des deux), mais elles nous sont de peu d'utilité pour l'évaluation de l'efficacité des méthodes de correction du poids de non-réponse dans la réduction du biais de non-réponse du panel.

## 5. DISCUSSION

On fait grand usage des poids de non-réponse afin de compenser la non-réponse des unités dans les enquêtes par sondages. L'exigence fondamentale de cette forme de pondération est la disponibilité des informations portant sur une ou plusieurs variables auxiliaires, tant pour les répondants que pour les non-répondants. Dans beaucoup d'enquêtes, ces informations ne sont disponibles que pour un petit nombre de variables auxiliaires (p. ex., UPE et strate à partir desquelles les unités ont été sélectionnées). Dans de telles enquêtes, les poids de non-réponse peuvent souvent prendre la forme de corrections de pondérations pour un ensemble de catégories, fondés sur la distribution croisée des variables auxiliaires.

Il existe toutefois des enquêtes dans lesquelles on peut compter sur un grand nombre de variables auxiliaires aux fins de la détermination possible de poids de non-réponse. Cette situation survient souvent lorsqu'on utilise un système de classement des dossiers administratifs en guise de base de sondage de l'enquête, l'ensemble des informations du système étant alors disponible pour la réalisation des corrections du poids de non-réponse. Tel est le cas également lorsque la collecte des données s'effectue en deux phases ou plus (c'est-à-dire, une entrevue initiale de présélection suivie d'une entrevue détaillée ou d'une forme quelconque de collecte de données à une date ultérieure) et lorsque les corrections du poids de non-réponse sont nécessaires pour les phases ultérieures. Dans ce dernier cas, les données des phases antérieures de collecte peuvent servir à la correction de la non-réponse au cours des phases ultérieures. C'est la même situation dans le cas des enquêtes par panel lorsqu'il devient nécessaire de procéder à des corrections pour la non-réponse au cours des vagues ultérieures de l'enquête, comme nous le décrivons dans le présent article.

Lorsqu'on dispose d'un grand nombre de variables auxiliaires pour chaque unité échantillonnée, deux choix doivent être faits. Premièrement, il faut choisir les variables auxiliaires à utiliser pour la correction. Deuxièmement il faut choisir la méthode de correction. Dans la présente étude, la méthode fondamentale retenue pour le choix des variables auxiliaires à utiliser pour la correction du poids de non-réponse a consisté à déterminer la série de variables qui représentait de bons prédicteurs de la non-réponse du panel. Compte tenu de la gamme de variables auxiliaires disponibles, la première

étape consistait à déterminer et à éliminer les variables qui n'étaient pas liées de près au taux de non-réponse du panel. Ensuite, des modèles de régression logistique utilisant les variables prédictives retenues à la présélection ont été examinés afin de déterminer les séries de variables à retenir aux fins de la correction de la pondération. Peu importe que le nombre de variables auxiliaires soit réduit à une taille raisonnable à l'aide de cette méthode ou d'une autre (p. ex., en utilisant l'algorithme CHAID) cette réduction risque d'être une première étape nécessaire lorsqu'on dispose d'un grand nombre de variables auxiliaires possibles.

Après la sélection du sous-ensemble de variables auxiliaires à utiliser, on peut choisir entre diverses méthodes pour la réalisation des corrections du poids de non-réponse. Nous avons examiné les corrections fondées sur les modèles de régression logistique, les modèles de recherches catégoriques et la méthode du quotient généralisée fondée sur l'échantillon. Les poids finals découlant de ces méthodes de correction montraient une très étroite corrélation entre eux et donnaient des estimations très semblables. Aucune des méthodes n'a produit des estimations meilleures que les autres pour la réduction du biais.

Les corrélations étroites observées entre les poids finals générés à l'aide des diverses méthodes de correction expérimentées peuvent s'expliquer en partie par les similitudes entre plusieurs de ces méthodes. Elles peuvent également s'expliquer en partie par la pondération finale de la stratification à posteriori qui a augmenté les corrélations entre les poids. Elle peut finalement s'expliquer en partie par l'absence d'effets d'interactions importants entre les variables auxiliaires. S'il y avait des effets d'interactions importants non inclus dans la modélisation logistique, on pourrait s'attendre à des différences plus grandes entre les poids obtenus par la méthode du quotient généralisée et par la méthode de correction logistique des taux de réponse prévus d'une part, et par la méthode CHAID, la méthode de correction des taux de réponses mixtes et la méthode de correction logistique en cellules groupées d'autre part. Ainsi, la similitude des poids produits à l'aide des diverses méthodes de pondération pour les données du SIP pourrait ne pas être aussi importante en d'autres circonstances.

Lorsqu'on utilise de nombreuses variables auxiliaires pour ajuster les poids, il arrive souvent que les résultats soient hautement variables, entraînant ainsi une sérieuse perte de précision des estimations de l'enquête. Ce problème ne s'est toutefois pas posé avec les méthodes que nous avons évaluées. La variabilité des poids obtenue avec l'ensemble des méthodes testées s'est avérée semblable, compte tenu du fait que des précautions raisonnables avaient été prises lors de la réalisation des corrections. Même si les résultats empiriques ne laissent pas voir de différences appréciables entre les estimations produites à l'aide des diverses méthodes de pondération testées et celles produites à l'aide des poids de l'enquête par panel SIP, les corrélations entre les poids ajustés des méthodes de rééchange et celui de l'enquête par panel SIP se sont avérées inférieures aux corrélations calculées entre les

Différences normalisées entre les estimations de l'enquête par panel SIPP de 1987 et les estimations repères

Méthode du quotient	Correction logarithmique					Estimations répères	Panel SIPP	Taux de réponse prévus	Taux de réponse mixtes	Cellules	CHAID 1	CHAID 2		
						AFDC	3.56	-1.58	-1.52	-1.43	-1.52	-1.44	-1.84	-1.57
						Coupons alimentaires	6.30	1.02	0.82	0.92	0.86	1.01	0.73	0.69
						Medicaid	6.97	-0.50	-0.47	-0.40	-0.53	-0.39	-0.70	-0.51
						SSI	1.65	0.05	0.11	0.08	-0.03	0.07	-0.15	0.09
						Sécurité sociale	15.14	-0.38	-0.46	-0.42	-0.44	-0.44	-0.50	-0.54
						Taux de pauvreté	14.46	-2.77	-2.64	-2.57	-2.67	-2.63	-2.74	-2.74
						Revenu médian	21,550	2.05	2.01	1.89	2.30	2.30	2.29	2.09
						Employés	61.60	2.42	1.60	1.56	1.76	1.72	1.95	1.73
						Chômeurs	4.52	-4.93	-4.59	-4.59	-4.76	-4.90	-4.78	-4.60
						Inactifs	33.88	-0.42	0.28	0.32	0.18	0.28	-0.01	0.15
Autres estimations répères														
						AFDC - Juin 1987	4.28	-2.55	-2.66	-2.49	-2.59	-2.65	-3.14	-2.71
						AFDC - Janvier 1989	4.24	-5.71	-5.62	-5.49	-5.63	-5.51	-6.10	-5.70
						Coupons alimentaires - Juin 1987	7.35	0.27	-0.31	-0.16	-0.04	0.11	-0.50	-0.48
						Coupons alimentaires - Janvier 1989	7.29	-2.04	-2.32	-2.17	-2.26	-2.06	-2.44	-2.50
						SSI - Juin 1987	1.68	0.00	0.13	0.08	-0.03	0.08	-0.20	0.11
						SSI - Janvier 1989	1.74	-0.57	-0.48	-0.53	-0.67	-0.54	-0.84	-0.50
						Salaires annuels 1987	2.22	-16.12	-15.94	-16.38	-15.66	-15.61	-15.60	-15.78
						Marités en 1987	1.86	-5.11	-4.93	-4.98	-5.11	-5.10	-5.07	-4.95
						Divorcés en 1987	0.90	-7.15	-7.37	-7.36	-7.40	-7.32	-7.20	-7.40
						Changement d'adresse en 1987	17.99	-11.49	-10.50	-10.51	-10.80	-10.42	-10.40	-10.49

Il importe avant tout de noter, dans le tableau 4, la similitude entre les estimations calculées avec les diverses méthodes de pondération à partir des données du SIPP de 1987. Les pourcentages estimatifs du tableau 4 sont en fait présentés avec une précision à une décimale ne laisserait souvent voir aucune différence entre les valeurs obtenues. La différence la plus importante s'observe dans le cas du pourcentage de personnes détendant un emploi en janvier 1989, où l'estimation utilisant le poids de l'enquête SIPP est de 62,7%, alors que celle obtenue à l'aide du poids de la méthode logarithmique à taux de réponse mixtes est de 62,3%. Même cette différence, qui est la plus grande observée, est relativement petite, surtout si on songe que l'erreur-type estimée correspondante n'est que de 0,3%.

Lorsqu'on compare les estimations de l'enquête par panel SIPP de 1987 aux estimations externes provenant de l'enquête par panel SIPP de 1989 on d'autres sources, certaines des différences observées sont plus grandes et beaucoup plus importantes. Pour les examiner d'une manière plus détaillée, nous avons calculé les différences normalisées entre les estimations des méthodes de rechange

La portion supérieure du tableau 5 présente les différences normalisées obtenues lorsque les données de l'enquête SIPP de 1989 sont utilisées pour produire l'estimation externe. Les différences normalisées pour la plupart des estimations sont inférieures à 2.0 en valeur absolue, ce qui signifie qu'elles peuvent être entièrement attribuables à des erreurs d'échantillonnage. Toutefois, les différences normalisées dans le cas du pourcentage de chômeurs et du taux de pauvreté sont supérieures à 2.0 et donc hautement significatives. Ainsi, les méthodes de rechange pour la correction de la pondération ne permettent pas d'aligner les estimations de l'enquête de 1987 sur celles de l'enquête de 1989 pour toutes les caractéristiques examinées. La portion inférieure du tableau 5 montre les différences de la différence.

normalisées obtenues lorsque d'autres estimations répétées sont utilisées. Ces différences normalisées sont généralement grandes et, dans beaucoup de cas, très grandes. Seules quelques-unes sont inférieures à 2,0, et beaucoup sont supérieures à 10,0. Etant donné les différences normalisées beaucoup plus petites notées dans la portion supérieure du tableau 5 pour des statistiques similaires, il



Estimations pour la population totale des données de l'enquête par panel SIPP de 1987 avec les méthodes de pondération de rachat, et estimations provenant d'autres sources

		Correction logistique							
		CHAID 1 CHAID 2							
		Méthode du quotient			Taux de réponse prévus mixtes	Taux de réponse Cellules groupées			
Panel Réfé-rence SIPP 1989	4.28 <sup>1</sup> 4.24 <sup>2</sup> 3.56	AFDC - Juin 1987 AFDC - Janvier 1989 AFDC - Année 1987	3.73 3.10 4.85	3.70 3.12 4.78	3.74 3.14 4.82	3.72 3.12 4.81	3.71 3.14 4.80	3.60 3.02 4.69	3.69 3.10 4.78
	7.35 <sup>3</sup> 7.29 <sup>4</sup> 6.30	Coupons alimentaires - Juin 1987 Coupons alimentaires - Janvier 1989 Coupons alimentaires - Année 1987	7.43 6.71 10.30	7.26 6.63 10.11	7.30 6.67 10.16	7.34 6.64 10.18	7.38 6.70 10.24	7.20 6.59 10.05	7.21 6.58 10.06
	6.97	Medicaid - Janvier 1989 Medicaid - Année 1987	6.77 9.21	6.78 9.21	6.81 9.24	6.75 9.21	6.81 9.25	6.68 9.09	6.76 9.21
	1.68 <sup>3</sup> 1.74 <sup>3</sup> 1.65	SSI - Juin 1987 SSI - Janvier 1989 SSI - Année 1987	1.68 1.65 1.80	1.70 1.67 1.82	1.69 1.66 1.82	1.67 1.64 1.80	1.69 1.66 1.82	1.65 1.61 1.78	1.69 1.66 1.82
	14.46	Taux de pauvreté - Juin 1987 Taux de pauvreté - Janvier 1989 Au-dessus du seuil de pauvreté 1987/1988	10.88 12.91 2.25	10.75 12.98 2.31	10.79 13.02 2.32	10.76 12.97 2.30	10.79 12.99 2.29	10.69 12.91 2.32	10.74 12.93 2.31
	2.22 <sup>4</sup> 2.550	Nombre moyen de mois sans assurance-maladie - 1987 Revenu médian du ménage - Janvier 1989 Salaires annuels 1987 (billions \$)	1.66 2,601 1.93	1.69 2,600 1.94	1.70 2,597 1.93	1.67 2,607 1.94	1.67 2,607 1.94	1.69 2,607 1.94	1.69 2,602 1.94
	61.60 4.52 33.88	Employés - Janvier 1989 Chômeurs - Janvier 1989 Inactifs - Janvier 1989	62.74 3.57 33.69	62.36 3.64 34.01	62.34 3.63 34.03	62.43 3.60 33.96	62.42 3.58 34.01	62.52 3.60 33.88	62.42 3.63 33.95
Maris en 1987	1.39	Divorcés en 1987	0.51	1.41	0.50	1.39	0.50	1.39	1.41
Changement d'adresse en 1987	12.88								
	17.99 <sup>6</sup> 1.86 <sup>5</sup> 0.90 <sup>6</sup>								

- 6 U.S. Bureau of the Census, Current Population Reports, Population Characteristics, P-20, No 473.
- 5 National Center for Health Statistics: Vital Statistics of the U.S., 1987, Volume III, Marriage and Divorce, DHHS Pub. No (PHS) 91-1103.
- 4 U.S. Bureau of the Census, Current Population Reports, Consumer Income, P-60, No 174.
- 3 USDA Food and Nutrition Service, données inédites.
- 2 Social Security Bulletin, Volume 52, No 3.
- 1 Social Security Bulletin, Volume 51, No 7.

La méthode de stratification a posteriori retenue était équivalente à celle utilisée dans le cadre de l'enquête SIPP, sauf que dans ce dernier cas, la stratification s'effectuait par groupes de renouvellement (dans le cas des méthodes de pondération de rechange, la stratification a posteriori s'effectuait sur l'ensemble des groupes de renouvellement combinés). Cette différence ne devrait pas avoir un effet appréciable. Après la stratification a posteriori, les six séries de poids finals et les poids du panel de l'enquête SIPP s'additionnent pour donner les mêmes totaux de contrôle.

Pour comparer les poids finals obtenus à l'aide des six méthodes de correction, d'abord entre eux puis aux poids de l'enquête par panel SIPP, nous avons calculé les corrélations entre les poids ainsi que la mesure de la variabilité utilisée antérieurement,  $(1 + CV^2)$ . Les résultats sont présentés au tableau 3. Les estimations de la variabilité due à la pondération  $(1 + CV^2)$  laissent constater des augmentations semblables oscillant entre 8% et 10% pour les variances des estimations de l'enquête de chacune des méthodes de pondération. Celles établies entre les séries de poids finals de l'enquête par panel sont toutes égales ou supérieures à 0,85. Si on compare ces corrélations à celles présentées au tableau 2, il paraît évident que les corrélations entre les poids finals sont sensiblement plus élevées que celles établies entre les poids de non-réponse du panel. Les corrélations entre le poids de l'enquête par panel SIPP et les poids finals des méthodes de rechange sont régulièrement inférieures à toutes les autres, ce qui s'explique probablement du fait que les variables utilisées aux fins des corrections du poids de non-réponse pour le premier différaient de ceux utilisés pour les seconds. Les variables utilisées dans les méthodes de rechange et qui n'étaient pas utilisées dans la pondération de l'enquête par panel SIPP étaient l'âge, le lien de parenté avec la personne de référence, le nombre d'items imputés, la catégorie d'emploi et l'admissibilité au programme de coupons alimementaires. La taille du ménage était la seule variable autre que les renseignements sur la MSA (exclus pour des raisons de confidentialité) prise en compte dans les poids de l'enquête par panel SIPP et ignorée dans les méthodes de rechange parce que son lien avec les taux de réponse n'avait pas été jugé significatif.

#### 4. COMPARAISON DES ESTIMATIONS OBTENUES À L'AIDE DES DIVERSES MÉTHODES DE PONDERATION

Dans la section précédente, nous avons décrit l'élaboration des séries de poids de rechange qui peuvent être utilisées pour l'analyse des données de l'enquête par panel SIPP. Toutes les méthodes de pondération de rechange examinées incorporent des corrections pour tenir compte des probabilités de sélection inégales, de la non-réponse de la vague initiale, de la non-réponse du panel et de la stratification a posteriori dans les totaux de contrôle externes. Dans la présente section, nous comparons les estimations

obtenues à l'aide des méthodes de pondération de rechange, d'abord l'une à l'autre puis aux estimations correspondantes obtenues avec les poids de l'enquête SIPP. En outre, le cas échéant, les diverses estimations de l'enquête sont également comparées aux estimations externes provenant d'autres sources. Certaines des estimations externes sont des estimations de référence obtenues à partir des données administratives de l'Étude de la population actuelle. D'autres estimations externes proviennent de la vague 1 de l'enquête par panel SIPP de 1989. Les données recueillies à la vague 7 de l'enquête par panel SIPP de 1987 portent sur la même période que celles recueillies à la vague 1 de l'enquête SIPP de 1989, ce qui signifie que les estimations provenant de ces deux sources de données devraient être comparables.

Pour faire les comparaisons avec les estimations de référence, il convient de reconnaître que toute différence relevée peut être due à une variété de facteurs parmi lesquels la non-réponse du panel ne constitue qu'une possibilité. Par exemple, les erreurs de réponses et les différences entre les définitions peuvent expliquer les différences observées entre les estimations de l'enquête SIPP et les estimations de référence. Il convient donc d'utiliser de prudence lors des comparaisons avec les estimations de référence. Comme les estimations de l'enquête par panel SIPP de 1989 sont fondées sur les données de la vague 1, elles sont à l'abri du problème de la non-réponse du panel. Ainsi, les différences observées entre les estimations obtenues à partir des enquêtes par panel SIPP de 1987 et de 1989 sont peut-être celles qui risquent le plus d'être causées par l'incapacité des corrections du poids de non-réponse du panel à compenser entièrement le biais de non-réponse du panel. Toutefois, même dans ce cas, d'autres explications possibles telles que le conditionnement du panel pourraient contribuer aux différences observées (même si Pennell et Lepkowski (1992) montrent que le conditionnement du panel ne constitue pas un facteur important dans la plupart des estimations des enquêtes SIPP).

Nous présentons au tableau 4 une gamme d'estimations tirées des données d'enquêtes par panel SIPP de 1987 qui utilisent les poids de l'enquête SIPP ainsi que ceux obtenus à l'aide des six méthodes de pondération de rechange, ainsi que les estimations de référence correspondantes et les estimations tirées de l'enquête par panel SIPP de 1989, lorsqu'elles étaient disponibles. Les estimations sont données en pourcentages, sauf dans le cas des estimations du nombre moyen de mois sans assurance-maladie, du revenu médian des ménages et des salaires annuels. Les estimations de la situation d'emploi (pourcentage des personnes occupées, des chômeurs et des inactifs), qui correspondent à des sujets âgés de plus de 15 ans, et dans le cas des salaires annuels, qui correspondent à des personnes âgées de plus de 14 ans. Les estimations correspondent à trois périodes différentes: juin 1987, janvier 1989 et l'année civile 1987. Par exemple, les trois premières estimations du tableau 4 sont les pourcentages estimés de personnes participant au programme AFDC (Aid to Families with Dependent Children)



Puisque le taux de réponse global pondéré du panel est de 0,794, la correction moyenne globale de la non-réponse serait de  $1/(0,794) = 1,26$  si la même correction était utilisée pour toutes les personnes. Les corrections de pondération moyennes pour les trois méthodes de correction de la pondération qui utilisent l'inverse des taux de réponse des cellules (méthode de correction logarithique en cellules groupées, CHAID 1 et CHAID 2) sont nécessairement égales à la correction globale de la non-réponse de 1,26. Les corrections de pondération moyennes des autres méthodes ne diffèrent que de façon minime de la correction moyenne globale du poids de non-réponse.

### 3.2 Poids finals du panel

3  
 steriori et mesures de l'augmentation de la variance

		Correction logistique					
Méthode du quotient		Panel SIPP		Taux de réponse prévus		Taux de réponse mixtes	
		CHAI 1	CHAI 2	Cellules groupées	Cellules groupées	Cellules groupées	Cellules groupées
Panel SIPP	Correction logistique	1.00	0.75	0.74	0.75	0.71	0.68
CHAI 1	taux de réponse prévus	1.00	0.99	0.91	0.90	0.86	0.98
CHAI 1	taux de réponse mixtes		1.00	0.91	0.90	0.86	0.97
CHAI 1	cellules groupées			1.00	0.89	0.85	0.93
CHAI 2					1.00	0.94	0.91
Méthode du quotient						1.00	1.00
1 + CV <sup>2</sup>		1.08	1.09	1.09	1.08	1.09	1.10

possession d'instruments financiers) en plus du sexe. Ce modèle a donné 99 cellules de correction de la non-réponse. La correction de la non-réponse fondé sur ce modèle a reçu le nom de CHAID 1. Le deuxième modèle CHAID comprendait les 13 variables de prévision du modèle de régression logistique énumérées au tableau 1. Ce modèle a donné 142 cellules de correction de la non-réponse. La correction de la non-réponse pour ce modèle a reçu le nom de CHAID 2.

**3.1.2 Corrections fondées sur la méthode du quotient généralisée**

La troisième catégorie de méthodes de correction du poids de non-réponse du panel est fondée sur la méthode du quotient généralisée. À la différence des autres catégories de méthodes, l'élaboration des cellules de correction de la non-réponse n'a pas été fondée sur la classification croisée des variables prédictives. Nous avons plutôt utilisé la méthode du quotient pour forcer l'égalité entre les distributions marginales des répondants du panel pour chaque variable de prévision. Nous avons utilisé les 10 variables prédictives du modèle de régression logistique réduit pour définir les distributions marginales. Ainsi, le problème de correction comportait 10 dimensions: une pour chaque variable de prévision.

La méthode du quotient exige une modification des poids originaux pour satisfaire à certaines contraintes marginales tout en minimisant la distance qui sépare les poids originaux des poids ajustés. Deville et Särndal (1992)

Tableau 2

Distribution des paramètres d'ajustement du poids de non-réponse du panel					
Corrélations	Correction logistique				
	Taux de réponse prévus	Taux de réponse mixtes	Cellules groupées	CHAID 1	CHAID 2
Correction logistique	1.26	1.04	1.20	4.28	1.02
Taux de réponse prévus	1.26	1.00	1.20	4.28	1.03
Taux de réponse mixtes	1.26	1.00	1.20	4.28	1.02
Cellules groupées	1.26	1.02	1.22	3.49	1.03
CHAID 1	1.26	1.01	1.19	13.93	1.04
CHAID 2	1.26	0.91	1.23	2.51	1.02
Méthode du quotient	1.26	0.96	0.73	0.63	0.95
CHAID 2	1.00	0.73	0.72	0.63	0.90
CHAID 1	1.00	0.73	0.69	0.81	0.73
Cellules groupées	1.00	0.73	0.58	1.00	0.75
Taux de réponse prévus	0.96	0.73	0.63	0.63	0.95
Taux de réponse mixtes	1.00	0.73	0.63	0.63	0.90
Cellules groupées	1.00	0.73	0.63	0.63	0.90
CHAID 1	1.00	0.73	0.63	0.63	0.90
CHAID 2	1.00	0.73	0.63	0.63	0.90
Méthode du quotient	1.00	0.73	0.63	0.63	0.90

décrivent certaines fonctions de distance auxquelles on peut recourir et obtiennent les méthodes de correction correspondantes. L'algorithme de la méthode du quotient de Deming et Stephan (1942), qui utilise implicitement une fonction de distance conduisant à une solution multiplicative, constitue une des formes de correction fondée sur la méthode du quotient généralisée. Nous avons utilisé le logiciel CALMAR décrit par Deville, Särndal et Sautory (1993) pour calculer les corrections. Trois fonctions de distance différentes ont été examinées: la méthode multiplicative, la méthode linéaire et la méthode multiplicative tronquée. Les corrections correspondantes à chacune des trois fonctions de distance se sont avérées presque identiques. Ce résultat empirique est conforme aux résultats obtenus par Deville et Särndal (1992) qui montrent que les estimateurs qui utilisent des poids générés avec des fonctions de distance différentes sont asymptotiquement équivalents si les fonctions de distance répondent à certaines conditions de lissage. Les trois fonctions de distance utilisées dans la présente recherche répondent à ces conditions. Comme les corrections avec chacune des trois méthodes étaient presque identiques, la méthode multiplicative a seule été retenue aux fins des évaluations ultérieures. Nous donnons à cette méthode le nom de correction par la méthode du quotient.

**3.1.3 Distributions des corrections de la non-réponse**

Les corrections correspondant à chacune des six méthodes décrites plus haut ont été calculées pour les données de l'enquête SIPP de 1987. Nous présentons au tableau 2 les distributions des corrections du poids de non-réponse ainsi obtenues. Seules les corrections y sont indiquées; il n'est pas question des poids tirés de ces corrections ni des poids de la vague 1. Le tableau 2 est divisé en



du taux de réponse observé (pondéré) dans chaque groupe de cellules correspond à la correction de la non-réponse du panel pour ce groupe. La correction de la non-réponse du panel est ensuite multipliée par le poids de la vague 1 afin de créer le poids corrigé de la non-réponse. Le poids de la vague 1 comprend une correction pour la non-réponse de la vague 1, mais il ne comprend pas celle de stratification a posteriori pour cette vague.

Dans la présente section, nous examinons diverses méthodes de correction des poids de non-réponse du panel. Nous abordons ci-après chacune de ces méthodes. Nous décrivons en détail les procédures d'élaboration de correction des poids et les propriétés statistiques importantes des corrections.

### 3.1 Corrections fondées sur les modèles logistiques

La première série de corrections de la pondération que nous examinons s'inspire directement du modèle de régression logistique décrit dans la section précédente. Cette correction de la pondération de la non-réponse du panel, appelée «correction logistique des taux de réponse prévus», a été calculée en utilisant l'inverse des taux de réponses prévus du modèle de régression logistique des effets majeurs réduits pour chacune des cellules de la classification croisée des 10 variables prédictives de ce modèle.

Comme les paramètres utilisés pour le calcul des taux de réponse prévus sont estimés avec un modèle des effets majeurs à partir des cellules de la classification croisée de la petite taille des cellules, la classification croisée de toutes les variables ne pose pas de difficultés. Toutefois, cet avantage est obtenu au prix d'une dépendance complète vis-à-vis de la validité du modèle des effets majeurs; autrement dit, nous devons presumer qu'il n'existe pas, entre les variables, d'interaction dont nous devrions tenir compte. Un des moyens de réduire notre dépendance vis-à-vis du modèle des effets majeurs consiste à fonder les corrections sur les taux de réponses observés dans les cellules dont les échantillons sont suffisamment grands pour assurer la stabilité des taux de réponse, et de fonder les corrections sur les taux de réponse prévus dans les autres cellules. La deuxième méthode de correction fondée sur la régression logistique utilise cette stratégie mixte. Dans les cellules qui contiennent 25 personnes échantillonnées ou plus, la correction de la non-réponse correspond à l'inverse du taux de réponse observé dans la cellule. Dans les cellules qui contiennent moins de 25 personnes échantillonnées, la correction de la non-réponse correspond à l'inverse du taux de réponse prévu pour chaque cellule. On donne à cette méthode le nom de «correction logistique des taux de réponse mixtes».

Une troisième méthode de correction du poids de non-réponse fondée sur les modèles logistiques rappelle les opérations utilisées dans le cadre de l'enquête SIPP. Les cellules initiales ont été définies par classification croisée

des 10 variables indépendantes utilisées dans la régression logistique. Les cellules ont ensuite été groupées jusqu'à ce que la taille de l'échantillon de chaque cellule dépasse 30, puis l'inverse du taux de réponse observé à l'intérieur de chaque groupe de cellules est devenu la correction de la non-réponse. On a pris soin de grouper les cellules qui présentaient des taux de réponses prévus comparables. On a donné à cette méthode de correction du poids de non-réponse le nom de «correction logistique en cellules groupées». Même si cette méthode ressemble à celle utilisée pour la correction de la non-réponse du panel dans le cadre de l'enquête SIPP, il existe certaines différences dans les variables utilisées pour définir les cellules et dans les méthodes utilisées pour combiner les petites cellules. Pour ces trois méthodes de correction du poids fondées sur le modèle de régression logistique, les taux de réponse observés et prévus ont été calculés à partir de dénombrements pondérés du nombre de cas, (au lieu d'utiliser des nombres non pondérés), en utilisant les poids corrigés de la non-réponse de la vague 1. En pratique toutefois, les corrections pondérées et non pondérées étaient presque identiques.

#### 3.1.1 Corrections fondées sur les modèles CHAID

La deuxième catégorie de méthodes de correction du poids de non-réponse du panel est fondée sur l'utilisation de l'algorithme de recherche par catégories CHAID pour la répartition de l'ensemble de données en cellules de correction. Il s'agit d'une manière générale d'assimiler des cellules de correction à des combinaisons de réponses aux variables prédictives qui présentent le plus grand pouvoir discriminatif en ce qui a trait aux taux de réponse du panel, à condition cependant que chaque cellule contienne un échantillon minimal de 25 personnes. La correction de la non-réponse du panel correspond à l'inverse du taux de réponse observé dans chaque cellule.

L'algorithme CHAID produit des cellules en séparant progressivement les séries de données en une structure arborescente. La séparation pour chaque nouvelle branche est effectuée en choisissant la variable qui maximise un critère  $\chi^2$ . Lorsque la séparation porte sur une variable polychotomique, elle risque d'intéresser plusieurs branches. Les tests de  $\chi^2$  sont modifiés en utilisant des corrections de type Bonferroni afin d'éviter que les variables ne soient choisies simplement parce qu'elles englobent un plus grand nombre de catégories. Le modèle CHAID est une version parmi d'autres du détecteur automatique d'interactions (DAI) mis au point pour les variables catégoriques. Kass (1980) expose la théorie qui sous-tend la technique CHAID. Lepkowski, Kalton et Kasprzyk (1989) et Kalton, Lepkowski et Lin (1985) ont utilisé une autre version de cette méthode pour la modélisation de la non-réponse dans les enquêtes SIPP.

En ce qui concerne l'analyse présente, nous avons examiné deux modèles CHAID incluant des séries différentes de variables prédictives. Le premier modèle comprenait les sept prédicteurs les plus importants du modèle de régression logistique (âge, lien de parenté avec la personne de référence, race du chef du ménage, statut d'occupation du logement, région du recensement, items imputés et

Estimations des paramètres du modèle de régression logistique

Prédicteurs		Estimations des paramètres	
<hr/>			
Coordonnée à l'origine		-0.465	
Âge ( $\chi^2 = 184.9, p < .0001$ ).		-0.179	
< 16		0.446	
16-24		0.187	
25-50		-0.056	
51-71		0.0	
> 71		-0.351	
Race ( $\chi^2 = 214.0, p < .0001$ ).		0.255	
Blanche		0.0	
Noire		-0.351	
Autre		0.0	
PPR ( $\chi^2 = 69.0, p < .0001$ ).		-0.251	
Membre de la famille		0.0	
Non-membre de la famille		-0.251	
Région de recensement ( $\chi^2 = 327.3, p < .0001$ ).		0.009	
Nouvelle-Angleterre		0.167	
Côte Atlantique moyenne		0.027	
Côte Atlantique sud		-0.231	
Centre sud-est		-0.396	
Centre nord		0.425	
Rochesuses/Centre sud-ouest		0.0	
Pacifique		-0.154	
Statut d'occupation du logement ( $\chi^2 = 207.2, p < .0001$ ).		0.331	
Propriétaire		0.0	
Locataire		-0.154	
Autre		0.0	
Items imputés ( $\chi^2 = 434.2, p < .0001$ ).		-0.626	
0		-0.244	
1		0.296	
2 à 3		0.0	
> 3		0.0	
Instruments financiers (obligations) ( $\chi^2 = 97.1, p < .0001$ ).		0.168	
Pas d'obligations		0.0	
Un certain nombre d'obligations		0.179	
Etat d'emploi ( $\chi^2 = 33.4, p < .0001$ ).		-0.179	
Non-chômeur		0.0	
Chômeur		-0.179	
Coupons alimentaires ( $\chi^2 = 39.3, p < .0001$ ).		-0.191	
Non-bénéficiaire		0.0	
Bénéficiaire		-0.191	
Catégorie d'emploi ( $\chi^2 = 31.4, p < .0001$ ).		0.100	
Affaires		0.103	
Autre		0.0	
Fonctionnaire		0.0	
Scolarité ( $\chi^2 = 12.8, p = .0003$ ).		-0.075	
Dixième ou onzième année		0.0	
Autre		-0.075	
Revenu du ménage ( $\chi^2 = 14.9, p = .0006$ ).		0.117	
Moins de \$1,200/mois		-0.088	
\$1,200 à \$8,000/mois		0.0	
Plus de \$8,000/mois		-0.088	
Sexe ( $\chi^2 = 10.3, p = .0013$ ).		0.047	
Homme		0.047	
Femme		0.0	
Rapport PPR / < 16 ans ( $\chi^2 = 10.1, p = .0015$ ).		0.096	
Membre de la famille, enfant		0.0	
Autre		0.0	

très étroitement de la propension à répondre risquent également d'être étroitement liés l'un à l'autre de telle sorte qu'un seul des deux pourrait suffire à la pondération. En conséquence, l'étape suivante de la sélection des prédicteurs de la non-réponse du panel a consisté à examiner quelles étaient les combinaisons d'items qui permettaient le mieux de prévoir la propension à répondre du panel.

Nous avons utilisé une méthode de régression logistique pour examiner les rapports conjoints de plusieurs items avec la propension à répondre du panel. Les modèles de régression ont été ajustés en utilisant les poids de la première vague qui tenaient compte des probabilités de sélection vagues et de la non-réponse à la première vague. Après avoir examiné un certain nombre de modèles possibles, nous avons retenu un modèle comportant 13 variables, d'effet majeur et un terme d'interaction, et offrant une représentation raisonnablement adéquate des données.

Nous présentons au tableau 1 les estimations des paramètres pour chaque niveau de chaque prédicteur de ce modèle, chacune accompagnée de la valeur correspondante de la statistique de Wald ( $\chi^2$ ). La valeur du paramètre du dernier niveau de chaque prédicteur (niveau repère) est fixée à zéro. Les estimations des paramètres des autres niveaux de chaque prédicteur représentent les différences dans la propension à répondre à partir du niveau repère. Comme on le constate à la lecture des statistiques de Wald, les prédicteurs jouent tous un rôle hautement significatif dans le modèle.

Ce modèle présente la particularité de ne contenir qu'un seul terme d'interaction, soit le rapport «lien de parenté avec la personne de référence/personne de moins de 16 ans». Toutes les autres interactions examinées présentaient des valeurs de  $\chi^2$  plus petites que cette dernière. Même le rapport «PPR / < 16 ans» avait une efficacité prédictive relativement faible. En fait, cette interaction ainsi que les trois derniers prédicteurs du tableau 1 (scolarité, revenu du ménage et sexe) ont été exclus de la plupart des opérations de pondération abordées plus loin à cause de leur puissance prédictive limitée de la propension à répondre du panel. Les méthodes de correction sont principalement fondées sur un modèle d'effets majeurs réduits comprenant les 10 premières variables prédictives énumérées au tableau 1.

### 3. AUTRES MÉTHODES DE CORRECTION DES POIDS

La méthode utilisée dans le cadre de l'enquête SIPP pour corriger la pondération de la non-réponse du panel est décrite par Chapman, Bailey et Kasprzyk (1986). Il s'agit essentiellement de former des cellules de correction de la non-réponse, puis de corriger les poids en utilisant l'inverse des taux de réponse de chacune des cellules. Les cellules sont formées par classification croisée des réponses tirées d'une série de variables de la vague 1 que l'on juge liées à la réponse du panel. Les petites cellules sont combinées de manière que la taille de l'échantillon obtenu dans chaque groupe de cellules soit de 30 ou plus. La réciproque



Aux fins de l'enquête SIPP, on procède chaque année à des entrevues auprès d'un échantillon national de ménages, et toutes les personnes âgées de quinze ans et plus vivant dans ces ménages à l'époque de la vague initiale deviennent les membres du panel qui fera l'objet du suivi subséquent. Jusqu'à maintenant, les panels de l'enquête SIPP ont eu une durée de vie de 32 mois, mais cette durée sera portée à quatre ans avec le panel de 1996. Les membres des panels sont interviewés tous les quatre mois. Les données recueillies portent sur le revenu, la participation aux programmes de soutien du revenu et d'autres facteurs qui peuvent influencer sur le revenu et la situation économique. On recueille également des données sur les enfants. Pour de plus amples informations sur le plan de sondage de l'enquête SIPP, voir Nelson, McMillen et Kasprzyk (1985) et Jabine, King et Petroni (1990).

L'étude qui fait l'objet du présent document a porté sur le panel de l'enquête SIPP de 1987. Les renseignements utilisés provenaient des bases de données à grande diffusion. Ce panel a débuté avec un échantillon d'environ 12,300 ménages, et le suivi comprenait en sept vagues successives de collecte de données. Le taux de non-réponse des ménages à la vague initiale était de 6,7% (Jabine et coll. 1990). Les ménages qui ont répondu à la vague initiale renfermaient 30,841 personnes, enfants compris. De ce total, 20,8% ont omis de fournir des données pour l'ensemble des vagues auxquelles elles étaient admissibles; c'étaient les non-répondants du panel.

Outre le choix des variables auxiliaires et l'étude des diverses méthodes d'utilisation de ces variables aux fins de la correction de la non-réponse du panel, nous avons procédé à une évaluation comparative des méthodes de estimations obtenues à l'aide de diverses méthodes de échange ont d'abord été comparées entre elles, puis comparées chacune à des estimations de référence. La partie finale du présent article résume les résultats obtenus et tire certaines conclusions sur l'efficacité des diverses méthodes de correction examinées. Rizzon, Kalton et Brick (1994) étudient ces questions d'une façon plus détaillée.

## 2. PRÉDICTEURS DE LA PROPENSION À RÉPONDRE

La première étape de la correction de la non-réponse du panel consiste à décider quels sont, parmi les nombreux items de la première vague de collecte de données, ceux qui devraient servir à la pondération. La présente section porte sur ce choix. Il s'agit de déterminer les items dont les réponses permettront de distinguer les participants en fonction de leur propension à répondre au cours des vagues subséquentes de l'enquête. Little (1986) appelle cette méthode une stratification de la propension à répondre; il montre que l'importation de biais des estimations peut être réduit en multipliant le poids de base par l'inverse de la probabilité de réponse d'un élément.

Dans l'enquête par panel SIPP de 1987, 58 des items de la vague initiale de collecte des données (vague 1) pouvaient servir de variables explicatives possibles de la

Puisque les 31 items retenus à l'issue de la présélection étaient au moins en partie liés à la non-réponse du panel, ils constituaient tous des variables utilisables aux fins de la pondération visant à réduire le biais de non-réponse du panel dans les estimations de l'enquête. Toutefois, la présélection a été imparfaite puisqu'elle n'a pas tenu compte des rapports existant entre les items, retenant ainsi plus de variables qu'il n'en fallait pour la correction de la non-réponse du panel. Par exemple, deux items qui dépendent l'un de l'autre ont été retenus, ce qui a entraîné une surpondération de la vague initiale d'une enquête par panel opérant moins à la vague initiale de la vague 1. D'autres études ont en effet démontré que les personnes qui coopèrent moins à la vague initiale de la vague 1. D'autres études ont en effet démontré que les personnes qui coopèrent moins à la vague initiale de la vague 1. D'autres études ont en effet démontré que les personnes qui coopèrent moins à la vague initiale de la vague 1.

Le dernier item (nombre d'items imputés) a été retenu et nombre d'items imputés à la vague 1. Le dernier item (nombre d'items imputés) a été retenu et nombre d'items imputés à la vague 1. Le dernier item (nombre d'items imputés) a été retenu et nombre d'items imputés à la vague 1.

La présélection nous a permis de réduire de 58 à 31 le nombre d'items de l'analyse de régression logistique. Voici quels ont été les items retenus: statut d'occupation du logement, logement social, type de ménage, région de recensement, scolarité du ménage, taille du ménage, revenu du ménage, possession d'instruments financiers (obligations), sexe, race, origine hispanique, lien de parenté avec la personne de référence (PPR), âge, état civil, type de famille, scolarité, état d'étudiant, état d'emploi, revenu personnel, emplois multiples, catégorie d'emploi, bénéficiaire des programmes Medicare, Medicaid, Women, Infants, and Children (WIC), Aid to Families with Dependent Children (AFDC) ou de coupons alimentaires, bénéficiaire d'un programme d'aide général, de la Sécurité sociale ou d'un autre programme de bien-être, état d'ancien combattant et nombre d'items imputés à la vague 1.

Le dernier item (nombre d'items imputés) a été retenu et nombre d'items imputés à la vague 1. Le dernier item (nombre d'items imputés) a été retenu et nombre d'items imputés à la vague 1. Le dernier item (nombre d'items imputés) a été retenu et nombre d'items imputés à la vague 1.



# Comparaison de quelques méthodes de correction de la non-réponse d'un panel

LOU RIZZO, GRAHAM KALTON et J. MICHAEL BRICK<sup>1</sup>

## RÉSUMÉ

Dans certaines enquêtes, on peut utiliser, pour la correction de la non-réponse, de nombreuses variables auxiliaires relatives aux répondants et aux non-répondants. On peut s'interroger sur le choix des variables auxiliaires à retenir aux fins de cette correction et sur la façon des les utiliser. Dans la présente recherche, nous examinons diverses méthodes de correction de la non-réponse fondées sur des modèles de régression logistique, sur des algorithmes de recherche par catégories et sur la méthode du quotient généralisée. Ces méthodes servent à la pondération de la non-réponse pour l'enquête Survey of Income and Program Participation (SIPP). Les estimations issues des diverses méthodes de correction sont comparées entre elles et également à des estimations de référence provenant d'autres sources.

MOTS CLÉS: Biais de non-réponse; enquêtes par panel; méthode du quotient généralisée; estimations de référence.

## 1. INTRODUCTION

On utilise généralement la pondération, dans l'analyse

de données d'enquêtes, afin de compenser les probabilités inégales de sélection des échantillons et pour la non-réponse de certaines unités, et de rendre les distributions de certaines variables de l'échantillon pondéré conformes à des distributions de population connues de ces variables (en compensant ainsi les carences dans la couverture et en améliorant la précision des estimations de l'enquête) (Kish 1992). Compte tenu de ces trois objectifs, la pondération est habituellement réalisée en trois étapes. Premièrement, un poids de base est calculé pour chaque élément échantilloné et correspond à l'inverse de la probabilité de sélection de chacun de ces éléments. Deuxièmement, les poids de base correspondent aux éléments répondants sont multipliés par un poids de non-réponse afin de corriger en fonction des non-répondants. Troisièmement, le poids ajusté est modifié pour faire en sorte que les distributions d'échantillons pondérées pour certaines variables soient conformes aux informations externes portant sur ces distributions.

Le présent article traite des méthodes de pondération qui cherchent à compenser la non-réponse des unités. Une des méthodes généralement utilisées pour le calcul des poids consiste à diviser l'échantillon total en une série de catégories de pondération fondées sur les informations recueillies tant sur les répondants que sur les non-répondants, puis d'augmenter les poids de base des répondants appartenant à une catégorie de pondération donnée afin de représenter les non-répondants de cette catégorie (Oh et Scheuren 1983; Kalton 1983). Dans beaucoup d'enquêtes, les informations disponibles sur les non-répondants, mises à part les unités primaires d'échantillonnage et la strate dans laquelle ils se trouvent, sont assez limitées. Dans de tels cas, le choix des catégories de pondération possibles est restreint et la méthode est donc relativement simple à appliquer.

Dans certaines enquêtes, toutefois, on possède une masse considérable d'informations sur les non-répondants. Ces informations peuvent provenir de la base de sondage (p. ex., échantillonnage d'employés à partir des dossiers du personnel) ou résulter d'un appariement des éléments échantillonnés et des données administratives. En outre, avec les enquêtes par panel et d'autres types d'enquêtes comportant plus d'une étape de collecte des données, les réponses recueillies lors des premières étapes fournissent une foule de renseignements sur les non-répondants que l'on peut utiliser aux étapes ultérieures.

La présente étude porte principalement sur les méthodes de correction de la non-réponse à utiliser lorsqu'on dispose de renseignements détaillés sur les caractéristiques des non-répondants. Dans une telle situation, le choix de la méthode à retenir dépendra des variables auxiliaires utilisées et de la façon dont on les utilisera.

Pour les fins de notre exposé, nous appliquerons diverses méthodes de pondération aux données d'une enquête particulière intitulée "Survey of Income and Program Participation" (SIPP). L'enquête SIPP est une enquête par panel réalisée en continu auprès des ménages par le U.S. Bureau of the Census. Les non-répondants d'un panel de l'enquête SIPP peuvent être répartis en deux groupes: ceux qui ne répondent pas à la première vague de collecte de données (non-répondants de la vague initiale), et ceux qui répondent à la vague initiale, mais qui ne répondent pas à la suivante ou aux vagues subséquentes du panel auxquels ils sont admissibles (non-répondants du panel). Dans ce dernier cas, la correction de la non-réponse du panel peut s'appuyer sur une masse importante de renseignements tirés de la vague initiale de collecte de données. Les méthodes de correction étudiées ici portent uniquement sur les non-répondants du panel. Ces méthodes modifient les poids des répondants du panel (c'est-à-dire, ceux qui fournissent des renseignements dans toutes les vagues où ils sont admissibles) et permettent de corriger en fonction des non-répondants.

<sup>1</sup> Lou Rizzo, Graham Kalton et J. Michael Brick, Westat Inc., 1650 Research Blvd., Rockville, MD 20850, U.S.A.





## BIBLIOGRAPHIE

- BHATTACHARYYA, G.K., et FRIES, A. (1986). On the inverse Gaussian multiple regression and model checking procedures. Dans *Reliability and Quality Control*, (A.P. Basu, Ed.). New York: North Holland, 86-100.
- CHAUVEY, Y.P. (1991). A study of ratio and product estimators under super population. *Communications in Statistics*, A, 20 (5 et 6), 1731-1746.
- CHOUDHRY, G.H., et RAO, J.N.K. (1988). Evaluation of small area estimators: an empirical study. Communication présentée à l'International Symposium on Small Area Statistics, New Orleans.
- CHHIKARA, R.S., et FOLKS, J.L. (1989). *The Inverse Gaussian Distribution*. New York: Marcel Dekker, Inc.
- DURBIN, J. (1959). A Note on the application of Quenouille's method of bias reduction to the estimation of ratio. *Biometrika*, 46, 477-480.
- FAY, R.E., et HERRIOT, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.
- FOLKS, J.L., et CHHIKARA, R.S. (1978). The inverse Gaussian distribution and its statistical applications-A review. *Journal of the Royal Statistical Society, Séries B*, 40, 263-275.
- FRIES, A., et BHATTACHARYYA, G.K. (1983). Analysis of two-factor experiments under an inverse Gaussian model. *Journal of the American Statistical Association*, 78, 820-826.
- GONZALES, M.E., et HOZA, C. (1978). Small area estimation with application to unemployment and housing estimates. *Journal of the American Statistical Association*, 73, 7-15.
- GHOSH, M., et RAO, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 55-76.
- HIDIROGLOU, M.A., et SÄRNDALE, C.-E. (1985). Etude empirique de quelques estimateurs de régression pour petits domaines. *Techniques d'enquête*, 6, 73-85.
- HOLT, D., SMITH, T.M.F., et TOMBERLIN, T.J. (1979). A model-based approach to estimation for small subgroups of a population. *Journal of the American Statistical Association*, 74, 405-410.
- IVENGAR, S., et PATWARDHAN, G. (1988). Recent developments in the inverse Gaussian distribution. Dans *Handbook of Statistics*, New York: Elsevier Science 479-490.
- JOHNSON, N. L., et KOTZ, S. (1970). *Continuous Univariate Distributions-I, Distributions in Statistics*. New York: Wiley.
- MACGIBBON, B., et TOMBERLIN, T.J. (1989). Estimation de proportions pour petites régions par des méthodes empiriques de Bayes. *Techniques d'enquêtes*, 15, 247-262.
- MICHAEL, J.R., SCHUCANY, W.R., et HASS, R.W. (1976). Generating random variables using transformations with multiple roots. *American Statistician*, 30(2), 88-90.
- PRAASAD, N.G.N., et RAO, J.N.K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, 85, 163-171.
- PURCELL, N.J., et KISH, L. (1980). Postcensal estimates for local areas (or domains). *Bulletin de l'Institut International de Statistique*, 48, 3-18.
- SÄRNDALE, C.-E. (1984). Design consistent versus model domain estimation: A conditional analysis. *Journal of the American Statistical Association*, 84, 266-275.
- SÄRNDALE, C.-E., et RÄBÄCK, G. (1983). Variance reduction and unbiasedness for small domain estimators. *Statistical Review*, 21, (Dissertations en l'honneur de T.E. Dalenius), 33-40.
- SCHABELE, W.L. (1979). A composite estimator for small area statistics. Dans *Synthetic Estimates for Small Areas*, NIDA Research Monograph 24, (J. Steinberg, Ed.). Rockville, MD: National Institute on Drug Abuse, 36-53.
- STATISTIQUE CANADA (1987). Fichier de microdonnées, Revenu des ménages et équipement des ménages (1987), Statistique Canada, Division des enquêtes-ménages.
- STROUD, T.W.F. (1987). Bayes and empirical Bayes approaches to small area estimation. Dans *Small Area Statistics*, (R. Platek, J.N.K. Rao, C.-E. Särndal et M.P. Singh, Eds.). New York: Wiley, 124-137.
- WHITMORE, G.A. (1983). A regression method for censored inverse Gaussian data. *La Revue Canadienne de Statistique*, 11, 305-315.



## Valeurs des paramètres servant à définir la population IG

## ANNEXE A

$d$	1	2	3	4	5
$10^6 \times \alpha_d$	3.1902855	2.8235779	1.5676078	.8056079	-.95350458

$d$	6	7	8	9	10
$10^6 \times \alpha_d$	-4.0661125	.49944356	.0061694263	-2.7414128	-1.1316622

$g$	1	2	3	4	5	6
$10^5 \times \beta_g$	1.0938451	.36781639	-.012707035	-.11561414	-.30936835	-1.023972

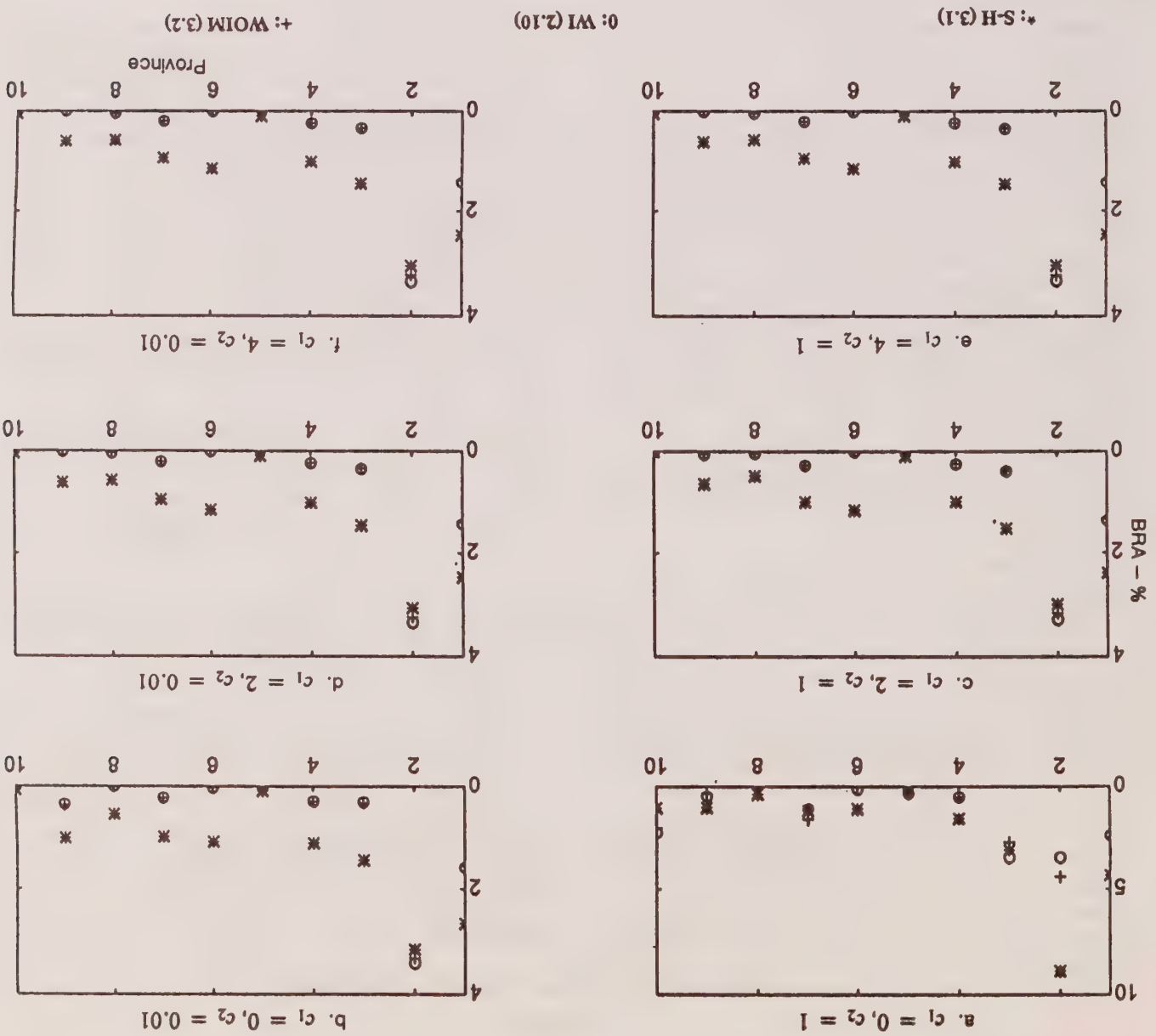
Valeurs de  $\theta_{dg}$ :

$d/g$	1	2	3	4	5	6
1	22,000.82	26,183.11	29,080.48	29,977.59	31,826.13	41,195.19
2	22,179.76	26,436.94	29,393.94	30,310.79	32,201.96	41,827.05
3	22,815.33	27,344.90	30,520.70	31,510.37	33,559.25	44,146.20
4	23,219.00	27,926.81	31,247.41	32,285.58	34,439.96	45,682.95
5	24,207.76	29,369.63	33,064.91	34,229.61	36,661.02	49,674.90
6	26,180.44	32,324.63	36,858.30	38,311.45	41,383.33	58,760.34
7	23,385.24	28,167.65	31,549.24	32,607.90	34,806.97	46,330.96
8	23,658.15	28,564.53	32,047.98	33,140.96	35,415.03	47,414.57
9	25,302.90	30,997.31	35,142.43	36,461.01	39,232.58	54,516.76
10	24,312.62	29,524.12	33,260.85	34,439.64	36,902.04	50,118.45

## ANNEXE B

Valeurs de la taille des cellules  $N_{dg}$ 

$d/g$	1	2	3	4	5	6	Total
1	627	360	277	84	215	110	1,673
2	285	212	198	72	68	83	918
3	597	483	616	148	204	231	2,279
4	729	397	568	151	239	219	2,303
5	1,372	761	1,216	202	473	511	4,535
6	1,177	888	1,795	517	707	800	5,884
7	639	432	673	165	236	222	2,367
8	850	512	888	264	349	297	3,160
9	700	699	1,350	385	696	572	4,401
10	456	540	1,083	342	393	407	3,221



#### 4. RÉSUMÉ ET CONCLUSIONS

Nous avons envisagé la généralisation de la méthodologie de l'analyse de la variance à une population gaussienne inverse à distribution asymétrique. Puis, nous avons étudié et appliqué les modèles sans interaction factorielle à l'estimation des paramètres régionaux d'une population finie. À partir de données d'enquêtes synthétiques grâce à une simulation des populations synthétiques grâce à une simulation de Monte Carlo. Nous avons montré ainsi que les estimateurs proposés donnent de bons résultats dans diverses conditions, si la population peut être considérée comme un échantillon aléatoire tiré d'une population à distribution gaussienne inverse. Cette méthode représente

un choix valable pour l'estimation des paramètres de données d'enquête dont la distribution présente une asymétrie positive.

## REMERCIEMENTS

Les professeurs Y.P. Chaubey et F. Nebebe remercient vivement le Conseil de recherches en sciences naturelles et en génie du Canada pour l'appui financier qu'il leur a accordé. Les auteurs remercient le rédacteur, les professeurs I.N.K. Rao et A.B. Sim, ainsi que les lecteurs extérieurs, pour leurs commentaires précieux.



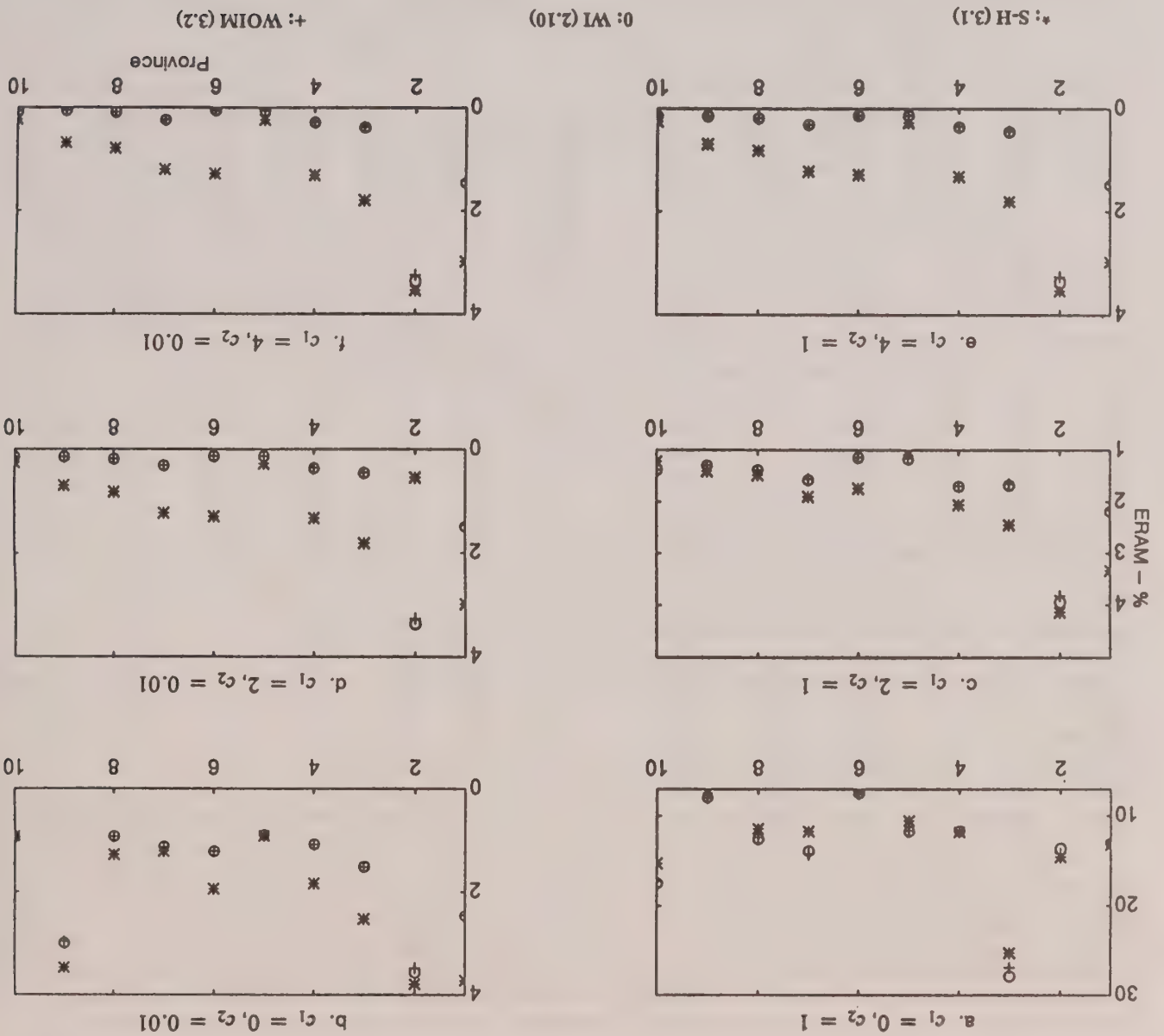


Figure 1. Erreur relative absolue moyenne de divers estimateurs pour un échantillon de 1%.

est faible, nous concluons que, pour le calcul d'estimations par modélisation, il est important de bien choisir le modèle de la moyenne quand le coefficient de variation est petit. Nous observons en outre que l'ERAM et le BRA des estimateurs  $\hat{t}^{dWOI}$  et  $\hat{t}^{dWOM}$  sont presque identiques, donc que la modification de l'estimateur représentée par (2.10) est inutile. Souignons, en revanche, que Hidiroglou et Särndal (1985) ont démontré que l'estimateur  $\hat{t}_{DS-H}$  est nettement supérieur à l'estimateur non corrigé correspondant proposé par Särndal (1984).

Certains ayant critiqué le fait que  $\hat{t}^{dWOI}$  et  $\hat{t}^{dWOM}$  dépendent du modèle, nous avançons les arguments qui suivent pour défendre le choix de ces estimateurs. La distribution gaussienne inverse peut prendre diverses formes et permet d'approcher la distribution lognormale, la distribution gamma, la distribution de Weibull et d'autres distributions à asymétrie positive. Donc, si l'on pense que la distribution de la caractéristique principale présente une asymétrie positive, la méthodologie que nous venons d'exposer est valable et utile.

Biais relatif absolu (en %) de divers estimateurs

Domaine	Échantillon de 1%				Échantillon de 5%				Échantillon de 1%				Échantillon de 5%			
	SH	WOI	WOIM	SH	WOI	WOIM	SH	WOI	WOIM	SH	WOI	WOIM	SH	WOI	WOIM	SH

c <sub>1</sub> = 0, c <sub>2</sub> = 1															
1	4,34	2,40	2,51	1,87	2,18	0,30	0,23	1,15	1,12	0,51	2,74	3,47	1,57	0,01	0,01
2	8,88	3,46	4,39	2,18	0,30	0,23	1,15	3,15	3,40	3,31	1,38	0,04	0,04	0,01	0,03
3	3,13	3,47	2,74	0,51	1,12	0,22	1,15	1,44	0,31	0,32	0,68	0,03	0,03	0,01	0,01
4	1,57	0,51	0,53	0,50	0,21	0,22	1,11	1,11	0,29	0,30	0,53	0,03	0,03	0,01	0,01
5	0,13	0,33	0,35	0,20	0,16	0,18	1,10	0,10	0,03	0,02	0,05	0,01	0,01	0,01	0,01
6	1,09	0,14	0,04	0,02	0,39	0,42	1,09	0,03	0,03	0,43	0,02	0,01	0,01	0,01	0,01
7	1,20	1,09	1,59	0,54	0,28	0,30	0,99	0,22	0,23	0,43	0,01	0,01	0,01	0,01	0,01
8	0,40	0,04	0,12	0,20	0,53	0,54	0,55	0,00	0,01	0,28	0,03	0,03	0,03	0,01	0,03
9	1,03	0,47	0,36	0,24	0,04	0,01	1,01	0,35	0,37	0,45	0,14	0,14	0,14	0,01	0,14
10	1,05	2,27	2,03	0,04	0,30	0,29	0,08	0,02	0,01	0,06	0,01	0,01	0,01	0,01	0,01
c <sub>1</sub> = 2, c <sub>2</sub> = 1															
1	2,40	1,37	1,33	1,13	0,01	0,01	2,47	1,43	1,39	1,15	0,01	0,01	0,03	0,00	0,00
2	3,00	3,28	3,16	1,33	0,02	0,04	3,06	3,34	3,24	1,36	0,03	0,03	0,03	0,00	0,00
3	1,53	0,39	0,38	0,70	0,04	0,04	1,46	0,35	0,34	0,65	0,01	0,01	0,01	0,00	0,00
4	1,00	0,25	0,25	0,53	0,04	0,04	1,01	0,23	0,23	0,49	0,00	0,00	0,00	0,00	0,00
5	0,10	0,02	0,03	0,04	0,00	0,01	0,10	0,01	0,02	0,04	0,00	0,00	0,00	0,00	0,00
6	1,16	0,01	0,01	0,47	0,02	0,02	1,15	0,01	0,00	0,46	0,00	0,00	0,00	0,00	0,00
7	1,00	0,27	0,27	0,42	0,00	0,00	0,95	0,21	0,21	0,41	0,00	0,00	0,00	0,00	0,00
8	0,48	0,04	0,04	0,25	0,01	0,01	0,57	0,04	0,04	0,26	0,00	0,00	0,00	0,00	0,00
9	0,64	0,06	0,05	0,27	0,02	0,02	0,61	0,01	0,00	0,26	0,00	0,00	0,00	0,00	0,00
10	0,01	0,02	0,02	0,02	0,00	0,00	0,06	0,01	0,01	0,03	0,00	0,00	0,00	0,00	0,00
c <sub>1</sub> = 4, c <sub>2</sub> = 1															
1	2,47	1,43	1,39	1,15	0,01	0,01	2,48	1,43	1,39	1,15	0,00	0,00	0,00	0,00	0,00
2	3,06	3,34	3,24	1,36	0,03	0,03	3,07	3,35	3,24	1,36	0,04	0,04	0,04	0,00	0,00
3	1,46	0,35	0,34	0,65	0,01	0,01	1,45	0,34	0,34	0,64	0,00	0,00	0,00	0,00	0,00
4	1,01	0,23	0,23	0,49	0,00	0,00	1,01	0,24	0,24	0,49	0,00	0,00	0,00	0,00	0,00
5	0,10	0,01	0,01	0,04	0,00	0,00	0,11	0,01	0,02	0,04	0,00	0,00	0,00	0,00	0,00
6	1,15	0,01	0,01	0,46	0,00	0,00	1,15	0,01	0,00	0,46	0,00	0,00	0,00	0,00	0,00
7	0,95	0,21	0,21	0,41	0,00	0,00	0,94	0,20	0,20	0,41	0,00	0,00	0,00	0,00	0,00
8	0,57	0,04	0,04	0,26	0,00	0,00	0,58	0,04	0,05	0,26	0,00	0,00	0,00	0,00	0,00
9	0,61	0,00	0,00	0,26	0,00	0,00	0,60	0,00	0,00	0,25	0,00	0,00	0,00	0,00	0,00
10	0,00	0,00	0,00	0,00	0,00	0,00	0,06	0,01	0,01	0,03	0,00	0,00	0,00	0,00	0,00
c <sub>1</sub> = 0, c <sub>2</sub> = .01															
1	2,66	1,58	1,54	1,22	0,03	0,03	2,66	1,58	1,54	1,22	0,03	0,03	0,03	0,00	0,00
2	3,15	3,40	3,31	1,38	0,04	0,04	3,15	3,40	3,31	1,38	0,04	0,04	0,04	0,00	0,00
3	1,44	0,31	0,32	0,68	0,01	0,01	1,44	0,31	0,32	0,68	0,01	0,01	0,01	0,00	0,00
4	1,11	0,29	0,30	0,53	0,03	0,03	1,11	0,29	0,30	0,53	0,03	0,03	0,03	0,00	0,00
5	0,10	0,03	0,02	0,05	0,01	0,01	0,10	0,03	0,02	0,05	0,01	0,01	0,01	0,00	0,00
6	1,09	0,03	0,03	0,43	0,02	0,02	1,09	0,03	0,03	0,43	0,02	0,02	0,02	0,01	0,01
7	0,99	0,22	0,23	0,43	0,01	0,01	0,99	0,22	0,23	0,43	0,01	0,01	0,01	0,01	0,01
8	0,55	0,00	0,00	0,28	0,03	0,03	0,55	0,00	0,00	0,28	0,03	0,03	0,03	0,01	0,03
9	1,01	0,35	0,37	0,45	0,14	0,14	1,01	0,35	0,37	0,45	0,14	0,14	0,14	0,01	0,14
10	0,08	0,02	0,01	0,06	0,01	0,01	0,08	0,02	0,01	0,06	0,01	0,01	0,01	0,01	0,01

Ici,  $t_d$  désigne un estimateur typique de  $t_d$  et  $t_{di}$  représente la valeur du  $i$ -ième échantillon de la simulation de Monte Carlo ( $i = 1, \dots, 1000$ ).

3.2 Analyse des résultats

Les valeurs de l'ERAM calculées conformément à (3.3) et celles du BRA calculées d'après (3.4) pour ces trois estimateurs et pour différentes tailles d'échantillon sont résumées aux tableaux 1 et 2, respectivement, pour certaines paires ( $c_1, c_2$ ). Les valeurs choisies pour  $c_1$  représentent des moyennes fortes (comme dans la population originale,  $c_1 = 0$ ), des moyennes modérées ( $c_1 = 2$ ), et des moyennes faibles ( $c_1 = 4$ ), tandis que celles choisies pour  $c_2$  représentent le paramètre de dispersion original ( $c_2 = 1$ ) et une autre valeur, plus faible ( $c_2 = .01$ ). Il est intéressant de noter que, si on augmente  $c_1$  d'une unité tandis que  $c_2$  demeure constant, le coefficient diminue d'un facteur 10.

La comparaison des valeurs de l'ERAM et du BRA obtenues pour les trois estimateurs révèle une diminution du biais ainsi que de l'erreur relative dans de nombreux cas, tant pour les échantillons de 1% que de 5%. On constate que les valeurs de l'ERAM et du BRA diminuent parallèlement à la valeur de la moyenne et à celle du paramètre de dispersion  $\sigma$ . Les diminutions sont appréciables, en particulier dans le cas de l'échantillon de 5% et (ou) quand la moyenne est nulle. On remarquera aussi que la diminution du biais est généralement plus importante que celle de l'erreur. Souignons également que, selon Johnson et Kotz (1970, pp. 141), pour une valeur constante de la moyenne, la distribution gaussienne inverse standardisée tend à devenir normale à mesure que le coefficient de variation tend vers zéro. Puisque l'augmentation de la valeur de l'ERAM et de celle du BRA est plus importante quand la valeur du coefficient de variation



$G, N_{dg}$  sont tirées de cette population (voir l'annexe B),  $D$  représentant le nombre de provinces (c.-à-d.  $D = 10$ ) et  $G$  représentant le nombre de groupes de niveau d'éducation (c.-à-d.  $G = 6$ ). Nous avons obtenu d'autres ensembles de valeurs de  $\theta_{dg}$  et de  $\sigma$  en considérant diverses combinaisons de  $(c_1, c_2)$ ;  $c_1 = 0(1)^4$  et de  $c_2 = 1, .25, .1, .01$ , où  $c_1$  est utilisé pour transformer  $\theta_{dg}$  en  $10^{-c_1\theta_{dg}}$  et  $c_2$ , pour transformer  $\sigma$  en  $c_2\sigma$ . Il convient de noter que quand  $c_1 = 0$  et  $c_2 = 1$ , on obtient les valeurs des paramètres de la population originale. En outre, les valeurs plus élevées de  $c_1$  indiquent des valeurs plus faibles des moyennes et celles de  $c_2$ , une valeur plus élevée du paramètre de dispersion.

Pour effectuer l'étude en simulation, nous avons commencé par produire, pour un ensemble donné de valeurs de  $\theta_{dg}$  et de  $\sigma$ , un échantillon aléatoire gaussien inverse au moyen de l'algorithme de Michael et coll. (1976), le nombre d'observations correspondant aux valeurs données à l'annexe B. Puis, nous sommes servis de cet échantillon aléatoire comme d'une population

Tableau 1

Erreur relative absolue moyenne (%) de divers estimateurs

Domaine	Echantillon de 1%				Echantillon de 5%				Echantillon de 1%				Echantillon de 5%			
	SH	WOI	WOIM	SH	SH	WOI	WOIM	SH	SH	WOI	WOIM	SH	SH	WOI	WOIM	SH
1	13.27	13.05	13.19	6.60	6.48	7.61	20.80	3.72	2.46	2.45	1.80	0.89	0.89	0.60	0.77	0.89
2	14.57	13.61	14.20	7.53	7.61	7.69	20.80	3.79	3.56	3.48	2.10	0.59	0.59	0.60	0.77	0.89
3	25.27	27.86	26.88	19.07	20.74	7.61	20.80	2.52	1.51	1.52	1.19	0.77	0.77	0.58	0.40	0.58
4	11.83	11.70	11.74	5.29	5.61	5.59	5.59	1.83	1.08	1.09	0.93	0.58	0.58	0.40	0.40	0.58
5	10.57	11.72	11.68	6.80	7.10	7.11	7.11	0.92	0.90	0.91	0.42	0.40	0.40	0.64	0.64	0.40
6	7.12	7.45	7.52	3.85	3.95	3.97	3.97	1.94	1.22	1.22	0.93	0.64	0.64	0.64	0.64	0.64
7	11.78	13.91	14.23	7.39	8.01	8.05	8.05	1.22	1.13	1.14	0.86	0.64	0.64	0.64	0.64	0.64
8	11.48	12.56	12.46	6.70	7.15	7.14	7.14	1.29	0.93	0.94	0.76	0.67	0.67	0.68	0.68	0.68
9	7.43	7.92	7.99	3.61	3.74	3.75	3.75	3.47	2.99	2.96	3.13	2.97	2.97	2.96	2.96	2.96
10	15.32	17.43	17.16	11.20	11.81	11.80	11.80	0.93	0.94	0.95	0.52	0.52	0.52	0.53	0.53	0.53
$c_1 = 0, c_2 = 1$																
1	3.34	2.18	2.15	1.66	0.79	0.78	0.78	2.99	1.48	1.47	1.47	0.08	0.08	0.13	0.07	0.08
2	4.14	3.94	3.82	2.14	1.07	1.06	1.06	0.54	3.37	3.27	1.86	0.14	0.14	0.08	0.08	0.13
3	2.44	1.67	1.65	1.17	0.71	0.70	0.70	1.81	0.45	0.44	0.87	0.07	0.07	0.07	0.07	0.07
4	2.05	1.70	1.69	0.98	0.70	0.70	0.70	1.32	0.36	0.35	0.66	0.07	0.07	0.05	0.05	0.07
5	1.08	1.17	1.16	0.50	0.51	0.51	0.51	0.27	0.13	0.13	0.11	0.05	0.05	0.05	0.05	0.05
6	1.74	1.14	1.14	0.78	0.52	0.52	0.52	1.29	0.13	0.13	0.13	0.05	0.05	0.05	0.05	0.05
7	1.90	1.57	1.56	0.91	0.72	0.72	0.72	1.22	0.31	0.31	0.56	0.07	0.07	0.06	0.06	0.07
8	1.48	1.38	1.38	0.70	0.60	0.60	0.60	0.81	0.18	0.18	0.38	0.06	0.06	0.06	0.06	0.06
9	1.41	1.30	1.29	0.67	0.59	0.58	0.58	0.69	0.14	0.14	0.30	0.06	0.06	0.06	0.06	0.06
10	1.22	1.38	1.38	0.56	0.59	0.59	0.59	0.26	0.15	0.15	0.10	0.06	0.06	0.06	0.06	0.06
$c_1 = 4, c_2 = 1$																
1	2.99	1.48	1.44	1.47	0.08	0.08	0.08	2.99	1.45	1.41	1.47	0.01	0.01	0.01	0.01	0.01
2	3.54	3.37	3.27	1.86	0.14	0.13	0.13	3.54	3.36	3.25	1.87	0.05	0.05	0.05	0.05	0.05
3	1.81	0.45	0.44	0.87	0.07	0.07	0.07	1.80	0.38	0.37	0.86	0.01	0.01	0.01	0.01	0.01
4	1.32	0.36	0.35	0.66	0.07	0.07	0.07	1.31	0.28	0.27	0.66	0.01	0.01	0.01	0.01	0.01
5	0.27	0.13	0.13	0.11	0.05	0.05	0.05	0.24	0.06	0.06	0.10	0.01	0.01	0.01	0.01	0.01
6	1.29	0.13	0.13	0.05	0.05	0.05	0.05	1.29	0.06	0.06	0.54	0.01	0.01	0.01	0.01	0.01
7	1.22	0.31	0.31	0.07	0.07	0.07	0.07	1.20	0.24	0.24	0.55	0.01	0.01	0.01	0.01	0.01
8	0.81	0.18	0.18	0.06	0.06	0.06	0.06	0.79	0.09	0.09	0.37	0.01	0.01	0.01	0.01	0.01
9	0.69	0.14	0.14	0.06	0.06	0.06	0.06	0.68	0.06	0.06	0.29	0.01	0.01	0.01	0.01	0.01
10	0.26	0.15	0.15	0.06	0.06	0.06	0.06	0.23	0.07	0.07	0.09	0.01	0.01	0.01	0.01	0.01
$c_1 = 4, c_2 = .01$																
1	2.99	1.48	1.44	1.47	0.08	0.08	0.08	2.99	1.45	1.41	1.47	0.01	0.01	0.01	0.01	0.01
2	3.54	3.37	3.27	1.86	0.14	0.13	0.13	3.54	3.36	3.25	1.87	0.05	0.05	0.05	0.05	0.05
3	1.81	0.45	0.44	0.87	0.07	0.07	0.07	1.80	0.38	0.37	0.86	0.01	0.01	0.01	0.01	0.01
4	1.32	0.36	0.35	0.66	0.07	0.07	0.07	1.31	0.28	0.27	0.66	0.01	0.01	0.01	0.01	0.01
5	0.27	0.13	0.13	0.11	0.05	0.05	0.05	0.24	0.06	0.06	0.10	0.01	0.01	0.01	0.01	0.01
6	1.29	0.13	0.13	0.05	0.05	0.05	0.05	1.29	0.06	0.06	0.54	0.01	0.01	0.01	0.01	0.01
7	1.22	0.31	0.31	0.07	0.07	0.07	0.07	1.20	0.24	0.24	0.55	0.01	0.01	0.01	0.01	0.01
8	0.81	0.18	0.18	0.06	0.06	0.06	0.06	0.79	0.09	0.09	0.37	0.01	0.01	0.01	0.01	0.01
9	0.69	0.14	0.14	0.06	0.06	0.06	0.06	0.68	0.06	0.06	0.29	0.01	0.01	0.01	0.01	0.01
10	0.26	0.15	0.15	0.06	0.06	0.06	0.06	0.23	0.07	0.07	0.09	0.01	0.01	0.01	0.01	0.01

finie dont nous avons tiré 1,000 échantillons aléatoires pour des fractions d'échantillonnage correspondant à 1% et à 5% avec remplacement. Nous avons, en fait, sélectionné plusieurs échantillons aléatoires et obtenu des résultats comparables à ceux exposés ici. Pour chaque échantillon, nous avons calculé les estimateurs des totaux  $t_{PS-H}$ ,  $t_{WOI}$  et  $t_{WOIM}$ . Nous avons choisi comme critères d'évaluation de la performance des estimateurs l'erreur relative absolue moyenne (ERAM) et le biais relatif absolu (BRA) définis comme suit:

$$\text{ERAM}(\hat{t}_d) = \frac{1}{1000} \sum_{i=1}^{1000} \frac{1000}{t_d} | \hat{t}_{di} - t_d | / t_d \quad (3.3)$$

$$\text{BRA}(\hat{t}_d) = \left| \frac{1}{1000} \sum_{i=1}^{1000} \frac{\hat{t}_{di} - t_d}{t_d} \right| \quad (3.4)$$

$$\mu_{j..} + \sum_{D=1}^D \alpha^D (y_{D.} - y_{D.}) + \sum_{G=1}^{G-1} \beta^G (y_{G.} - y_{G.}) = n_{..},$$

$$\mu(y_{D.} - y_{D.}) + \alpha^D y_{D.} + \sum_{j=1}^{j-1} \alpha^j y_{D.}$$

$$+ \sum_{G=1}^{G-1} \beta^G \{ (y_{Dg} - y_{Dg}) - (y_{Dg} - y_{Dg}) \} = n_{D.} - n_{D.},$$

$$\mu(y_{g.} - y_{g.}) + \sum_{D=1}^D \alpha^D \{ (y_{Dg} - y_{Dg}) - (y_{Dg} - y_{Dg}) \} + \sum_{G=1}^{G-1} \beta^G y_{j.G} + \sum_{j=1}^j \beta^j y_{j.G} = n_{g.} - n_{g.}, \quad (2.6)$$

où les totaux et les moyennes sont représentés par les expressions

$$y_{Dg} = \sum_k y_{Dgk}, y_{D.} = \sum_g y_{Dg}, y_{g.} = \sum_D y_{Dg}, \quad (2.7a)$$

$$n_{D.} = \sum_g n_{Dg}, n_{g.} = \sum_D n_{Dg}, n_{..} = \sum_g \sum_D n_{Dg}. \quad (2.7b)$$

Les solutions  $(\mu, \hat{\alpha}^D, \hat{\beta}^G)$ ,  $d = 1(1)D$ ,  $g = 1(1)G$ , qui fournissent le pseudo estimateur du maximum de vraisemblance, n'aboutissent pas nécessairement à des estimations de réponse non négatives, mais correspondent à l'estimateur du maximum de vraisemblance (EMV) correct quand  $n_{Dg} \rightarrow \infty$  (voir Fries et Bhattacharyya 1983), la probabilité étant égale à un. Nous pouvons donc tronquer les valeurs des estimateurs de réponse à zéro pour annuler les valeurs négatives.

Dans le cas du modèle  $IG(\theta, \sigma)$  avec interactions, le paramétrage habituel des effets d'interaction donne à penser que le modèle prend la forme

$$\theta_{-1}^{-1} = \mu + \alpha^D + \beta^g + \gamma^{Dg},$$

$$\sum \alpha^D = \sum \beta^g = \sum \gamma^{Dg} = 0, \quad (2.8)$$

où  $\gamma^{Dg}$  représente maintenant l'effet d'interaction quand on considère le  $d$ -ième domaine et le  $g$ -ième groupe. Dans ces conditions, nous pouvons obtenir les estimateurs des paramètres en appliquant la méthode décrite ci-dessus. Toutefois, puisque l'estimateur du maximum de vraisemblance (EMV) de  $\theta^{Dg}$  est  $y^{Dg}$ , et qu'il existe une relation de parité entre les paramètres du modèle réparé et en ce qui concerne  $(\mu, \alpha^D, \beta^g, \gamma^{Dg})$ , d'une part, et les paramètres originaux,  $\theta^{Dg}$ , d'autre part, il n'est pas nécessaire d'établir des formules explicites pour l'EMV de divers paramètres. Donc, selon l'équation (2.3), pour un modèle bifactoriel avec interactions, notre estimateur est

$$(2.9)$$

$$l_{LWI} = \sum_g N_{Dg} y_{Dg},$$

expression qui correspond à l'estimateur stratifié à poste-riori et ne présente donc aucun intérêt supplémentaire en ce qui concerne l'estimation des caractéristiques des petites régions. Dans le cas du modèle avec interactions, l'estimateur s'exprime par

$$l_{LWOI} = \sum_g N_{Dg} \theta_{Dg} + \sum_g N_{Dg} (y_{Dg} - \theta_{Dg}), \quad (2.10)$$

où  $\theta_{-1}^{-1} = \mu + \alpha^D + \beta^g$ , les estimateurs étant calculés à partir de (2.6) et de  $N_{Dg} = n_{Dg} N / n_{..}$ . Afin d'évaluer l'efficacité de cet estimateur, nous avons effectué une étude numérique dont les résultats sont exposés à la section suivante.

### 3. ÉTUDE NUMÉRIQUE DE L'ESTIMATEUR DE RÉGRESSION GAUSSIEN INVERSE

Nous présentons ici les résultats d'une étude en simulation effectuée pour évaluer la performance des estimateurs développés à la section précédente. À titre de référence, nous utilisons l'estimateur de régression corrigé proposé par Särndal et Hidiroglou (1989) dont voici l'expression:

$$l_{DS-H} = \sum_g N_{Dg} y_{g.} + \sum_g F_d N_{Dg} (y_{Dg} - y_{g.}), \quad (3.1)$$

où  $F_d = N_d / N_d$ , si  $N_d \geq N_d$ , sinon,  $F_d = N_d / N_d$ . Ici,  $N_d = n_d N / n_{..}$ . On peut obtenir une autre expression de cet estimateur, qui tient compte à la fois des effets de groupe et de domaine, en remplaçant  $y_{g.}$  par  $y_{g.} + y_{D.} - y_{..}$ , mais cette approche n'a pas été poursuivie ici. Il convient de noter qu'on ne peut calculer les estimateurs susmentionnés quand  $n_{Dg}$  est égal à zéro. Le cas échéant, on choisit simplement comme estimateurs les moyennes d'échantillons des domaines respectifs. Aux fins de comparaison, nous incluons également la version modifiée de  $l_{LWOI}$  que voici:

$$l_{LWOI} = \sum_g N_{Dg} \theta_{Dg} + \sum_g F_d N_{Dg} (y_{Dg} - \theta_{Dg}). \quad (3.2)$$

#### 3.1 Conception de l'étude en simulation

Nous nous sommes servis des données sur le revenu des ménages au Canada en 1986, tirées du fichier de micro-données sur le revenu des ménages et sur les équipements ménagers de Statistique Canada (1987), pour calculer la valeur des paramètres utilisés pour la simulation. Après avoir réparti les données sur le revenu des ménages en dix domaines, selon la province, et en six groupes, selon le niveau d'éducation, nous avons commencé par ajuster un modèle gaussien inverse représenté par l'équation (2.4). Puis, à partir des estimations des paramètres, nous avons calculé les paramètres réels de la superpopulation gaussienne inverse qui sont résumés à l'annexe A. Les valeurs de  $D$ ,



$\eta$  similaire à l'estimateur du paramètre de régression du modèle linéaire habituel (voir Särndal 1984) s'exprime par

$$\hat{\eta} = \left( \sum_{k \in S_d} \frac{\pi_k}{x_k x'_k y_k} \right)^{-1} \sum_{k \in S_d} \frac{\pi_k}{x_k}. \quad (2.2)$$

Cet estimateur est appelé pseudo estimateur du maximum de vraisemblance, car, puisqu'on le calcule par maximisation inconditionnelle de la fonction de vraisemblance, l'inégalité  $x'_k \eta > 0$  n'est pas nécessairement satisfaite pour toutes les valeurs de  $k$ . Par conséquent, à l'exemple de l'estimateur de régression corrigé de Särndal (1984), nous pouvons représenter l'estimateur du total  $t_d$  du  $d$ -ième domaine par la relation

$$t_{dIG} = \sum_{k \in U_d} y_k + \sum_{k \in S_d} \frac{\pi_k}{e_k} \quad (2.3)$$

où  $y_k = x'_k \eta$  et  $e_k = y_k - \hat{y}_k$ . Dans l'exposé qui suit, nous représentons la moyenne de la cellule  $(d, g)$  par  $\theta_{dg}$ , et nous envisageons un échantillon aléatoire simple, auquel cas  $\pi_k$  est constant. En premier lieu, nous examinons la prédiction des observations au moyen de l'équation (2.3), en nous fondant sur un modèle à effets additifs représenté par

$$\theta_{dg}^{-1} = \mu + \alpha_d + \beta_g, \quad \sum \alpha_d = \sum \beta_g = 0, \quad (2.4)$$

où  $\mu$ , les facteurs  $\alpha_d$  et  $\beta_g$  représentent respectivement l'effet global, les effets par domaine ou par rangée, et les effets par groupe ou par colonne. Dans le cas d'une distribution gaussienne inverse, il faut aussi que  $\theta_{dg} > 0$  pour toutes les cellules  $(d, g)$  et que  $\sigma > 0$ . Donc, les paramètres  $\mu$ ,  $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_D)$ ,  $\beta = (\beta_1, \beta_2, \dots, \beta_G)$ , ainsi que  $\sigma$  sont compris dans l'ensemble  $\Omega = \{(\mu, \alpha, \beta, \sigma) : \sum^D \alpha_d = 0, \sum^G \beta_g = 0; \mu + \alpha_d + \beta_g > 0, \forall (d, g); \sigma > 0\}$ . Dans ces conditions, on peut estimer les paramètres de prédiction par maximisation inconditionnelle de la fonction de vraisemblance. Compte tenu de la taille de la population et de celle des échantillons  $n_{dg}$ , et en se référant à (2.1) et à (2.3), la fonction logarithmique de vraisemblance des paramètres s'exprime par

$$\ell = -\frac{1}{2} \log \sigma \sum_d \sum_g n_{dg} - (2\sigma)^{-1} \sum_d \sum_g \sum_k y_{dgk} [\gamma_{dgk} (\mu + \alpha_d + \beta_g) - 1]^2. \quad (2.5)$$

En premier lieu, observons que les paramètres sont effectivement donnés par  $(\mu, \alpha_d, \beta_g, d = 1, 2, \dots, D - 1; g = 1, 2, \dots, G - 1)$ . Donc, par différentiation de l'équation susmentionnée par rapport à  $(\mu, \alpha_d, \beta_g, d = 1, 2, \dots, D - 1; g = 1, 2, \dots, G - 1)$  et en posant les dérivées partielles obtenues comme égales à zéro, on obtient les équations qui suivent pour les estimateurs  $(\hat{\mu}, \hat{\alpha}_d, \hat{\beta}_g, d = 1, 2, \dots, D - 1; g = 1, 2, \dots, G - 1)$ ,

bifactoriels au moyen d'un modèle gaussien inverse, est d'une très grande importance. Ces auteurs présentent le calcul d'estimations au moyen d'un modèle symétrique, sans interaction. Nous avons étendu cette approche au cas d'un modèle asymétrique, démarche essentielle à l'estimation des totaux ou des moyennes par domaine. À cet égard, on pourrait adapter la méthode générale de régression multiple de Bhattacharyya et Fries (1986), et Whitmore (1983), mais nous avons préféré la méthode directe. À la Section 2, nous précisons le modèle et présentons les estimateurs que nous proposons en vertu du modèle gaussien inverse. À la Section 3, nous effectuons une étude numérique en vue d'évaluer la performance de l'estimateur proposé grâce à une simulation de Monte Carlo. Enfin, à la Section 4, nous présentons un résumé et les conclusions.

## 2. LE MODÈLE DE RÉGRESSION GAUSSIEN INVERSE APPLIQUÉ AU CALCUL DES ESTIMATIONS RÉGIONALES

Supposons qu'on divise une population finie  $U$  en  $D$  domaines qui ne se chevauchent pas  $U_d, d = 1(1)D$ , où  $N_d$  représente la taille de  $U_d$ . La population est en outre divisée suivant une deuxième dimension, en  $G$  groupes qui ne se chevauchent pas  $U_{dg}, g = 1(1)G$ , où la taille de  $U_{dg}$  est représentée par  $N_{dg}$ . Le recoupement des domaines et des groupes produit  $DG$  cellules de population  $N$  par la relation  $N = \sum^D N_d = \sum^G N_{dg} = \sum^D \sum^G N_{dg}$ . Nous désirons estimer les totaux par domaine  $t_d = \sum U_d y_k$ , où  $y$  représente la variable caractéristique et  $y_k$ , l'observation pour la  $k$ -ième unité. Un échantillon  $s$  de taille  $n$  est tiré de  $U$  par simple échantillonnage aléatoire. Représentons par  $s_d, s_{dg}$  les parties de  $s$  qui tombent dans  $U_d$ ,  $U_{dg}$  et  $U_{dg}$ . Les tailles d'échantillon correspondantes sont représentées par  $n_d, n_g$  et  $n_{dg}$ , et respectivement.

### 2.1 Méthode de régression pour les données gaussiennes inverses

Nous conseillons au lecteur de consulter les revues détaillées des travaux récents sur la distribution gaussienne inverse publiées par Chhikara et Fols (1989) et par Iyengar et Patwardhan (1988). La densité de probabilité d'une variable gaussienne inverse dont les paramètres sont  $(\theta, \sigma), IG(\theta, \sigma)$  s'écrit

$$f(y; \theta, \sigma) = (2\pi\sigma)^{-1/2} y^{-3/2} \exp[-(2\sigma y)^{-1} (y\theta^{-1} - 1)^2]; \quad (2.1)$$

où  $y > 0, \theta > 0, \sigma > 0$ . La moyenne et la variance de cette distribution sont  $\theta$  et  $\theta^3 \sigma$ , respectivement. Bhattacharyya et Fries (1982) ont proposé un modèle linéaire réciproque pour  $\theta$ . Plus précisément, ils supposent que le modèle a la forme  $\theta_k^{-1} = x'_k \eta$ . Dans ces conditions, un estimateur de

# Estimation des caractéristiques des petites régions au moyen d'un modèle gaussien inverse

Y.P. CHAUBEY, F. NEBEBE et P.S. CHEN<sup>1</sup>

## RÉSUMÉ

Les auteurs examinent la méthodologie de l'analyse de la variance dans le cas d'une distribution gaussienne inverse et l'adaptent à l'estimation des paramètres régionaux de populations finies. Grâce à une étude de Monte Carlo, ils démontrent que le choix de ces estimateurs est défendable pour les données d'enquête étalées vers la droite, telles que celles sur le revenu ou sur le rendement d'un secteur particulier.

**MOTS CLÉS:** Interactions; gaussien inverse; Monte Carlo; estimations de régression; estimations synthétiques; estimateur de Särndal-Hidiroglou; modèle asymétrique.

## 1. INTRODUCTION

Un grand nombre de méthodes visant à résoudre la question des estimations pour des petites régions ont été publiées récemment, dont celles de Prasad et Rao (1990), Särndal et Hidiroglou (1989), Choudhry et Rao (1988), et Särndal (1984), ainsi que les études auxquelles se réfèrent ces auteurs, en particulier Särndal et Råbäck (1983), Fay et Herriot (1979), Shaible (1979), Holt, Smith et Tomberlin (1979), et Gonzalez et Hoza (1978), pour n'en nommer que quelques-unes. La nécessité d'obtenir des estimations régionales de plusieurs caractéristiques d'une population donnée a mené à l'élaboration de diverses méthodes pré-cieuses qui permettent de produire des estimations réalistes et suffisamment exactes pour des secteurs restreints et d'autres sous-groupes particuliers. Plusieurs techniques proposées par les auteurs susmentionnés sont, implicitement et (ou) explicitement, fondées sur un modèle et sur l'application de la distribution théorique normale standard. D'autres chercheurs se sont attaqués au calcul d'estimateurs pour les secteurs restreints selon une approche bayésienne théorique ou empirique, en recherchant un compromis entre la moyenne d'échantillon d'une région (dont la population est présumée normale) et un estimateur calculé par régression en fonction d'une ou de plusieurs covariables (voir, p. ex., Stroud 1987, MacGibbon et Tomberlin 1989). Pour une revue détaillée des progrès récents en ce qui concerne les estimations régionales, le lecteur devrait consulter Ghosh et Rao (1994).

L'application de la distribution théorique normale standard à l'analyse des plans d'expérience factoriels est parfois inappropriée quand les données sont produites à partir d'une population dont la distribution présente une asymétrie positive nette. Bien que la plupart des procédures d'inférence soient retragables sur le plan analytique, le cas de nombreuses applications pratiques. Par conséquent, dans de telles situations, une analyse fondée sur des distributions à asymétrie positive se justifie.

La présente étude vise à examiner les méthodes d'inférence applicables à des plans d'expérience bifactoriels, tant pour les petites régions. Hidiroglou et Särndal (1985) ont décrit une étude de Monte Carlo effectuée en choisissant un estimateur de régression corrigé comme compromis entre l'estimateur synthétique et l'estimateur de régression généralisé. Ces auteurs (Särndal et Hidiroglou 1989) ont aussi présenté d'autres comparaisons d'estimateurs fondées sur l'inférence conditionnelle. Fondamentalement, l'estimateur de régression généralisé est calculé d'après le modèle de régression d'une superpopulation, sans hypothèse quant à la distribution. Chaubey (1991), quant à lui, a examiné les modèles de superpopulation proposés par Durbin (1959) dans le cas d'une distribution gamma ou gaussienne inverse auxiliaire, auquel cas l'estimateur de régression généralisé s'avère le meilleur prédicteur linéaire non biaisé (voir Prasad et Rao 1990). En fait, le meilleur estimateur linéaire non biaisé pour l'ensemble de la population étant indépendant de la forme de la distribution de la variable caractéristique, cette méthode est préférable, compte tenu de la difficulté éventuelle à calculer les estimations du maximum de vraisemblance (EMV). Ayant observé que les distributions des superpopulations (transposées aux populations), peuvent être fort semblables aux distributions gaussiennes inverses de diverses populations, nous aimerions tirer parti de cette caractéristique.

Le recours à une distribution gaussienne inverse n'est pas un exercice futile. Cette distribution, qui a été appliquée avec succès dans de nombreuses situations (voir Folks et Chhikara 1978), est fort similaire à la distribution gamma, à la distribution lognormal et à la distribution de Weibull utilisées couramment pour la modélisation de variables aléatoires non négatives dont la distribution présente une asymétrie positive. Dans le présent article, nous étudions l'application du modèle gaussien inverse aux estimations régionales. L'approche de Fries et de Bhattacharyya (1983), lesquels discutent de l'analyse de plans d'expérience

<sup>1</sup> Y.P. Chaubey, Professeur, Department of Mathematics and Statistics; F. Nebbe, Professeur agrégé, Department of Decision Sciences & M.I.S.; et P.S. Chen, assistant à la recherche, Department of Finance, Concordia University, Montréal, Canada.





On parvient à l'approximation (A.1) en notant (i) que  $\hat{A}^{(g_l)} - \hat{A}$  est d'un ordre inférieur à  $\hat{A}$ , selon l'hypothèse que la contribution d'aucune grappe n'est disproportionnée, à mesure que le nombre de strates  $L$  augmente (lire Yung (1996) pour plus de précisions sur les conditions de régularité) et (ii) que  $[I + \hat{A}^{-1}(\hat{A}^{(g_l)} - \hat{A})]^{-1} \approx I - \hat{A}^{-1}(\hat{A}^{(g_l)} - \hat{A})$ .

De (A.1), on obtient

$$\hat{b}^{(g_l)} - \hat{b} = (\hat{A}^{-1} - \hat{A}^{-1} + \hat{A}^{-1})(\hat{b}^{(g_l)} - \hat{b} + \hat{b}) - \hat{A}^{-1}\hat{b}$$

$$\approx (\hat{A}^{-1} - \hat{A}^{-1})\hat{b} + \hat{A}^{-1}(\hat{b}^{(g_l)} - \hat{b})$$

$$\approx -\hat{A}^{-1}(\hat{A}^{(g_l)} - \hat{A})\hat{b} + \hat{A}^{-1}(\hat{b}^{(g_l)} - \hat{b}).$$

(A.2)

Étant donné (A.2), il s'ensuit que

$$Y^{(g_l)} - Y_r \approx (Y^{(g_l)} - Y) - (X^{(g_l)} - X)^T \hat{b}$$

$$-(X - X)^T (\hat{b}^{(g_l)} - \hat{b})$$

(A.3)

$$\approx \frac{1}{1} (e_g^* - e_{g_l}^*),$$

où  $e_{g_l}^* = \sum_k (n_g w_{gjk}^*) e_{gjk}$  et  $e_g^* = (1/n_g) \sum_j e_{gj}^*$ . On s'est servi des résultats ci-dessous pour parvenir à (A.3):

$$(Y^{(g_l)} - Y) - (X^{(g_l)} - X)^T \hat{b} = \frac{1}{1} (e_g - e_{g_l})$$

et

$$(X - X)^T (\hat{b}^{(g_l)} - \hat{b}) \approx$$

$$(X - X)^T \hat{A}^{-1} \left[ \frac{1}{1} (n_g - 1) (u_g - u_{g_l}) \right],$$

$$\text{où } e_{g_l} = \sum_k (n_g w_{gjk}) e_{gjk} \text{ et } u_{g_l} = \sum_k (n_g w_{gjk}) x_{gjk} e_{gjk}.$$

Par conséquent, de (A.3), on déduit que:

## BIBLIOGRAPHIE

- BEEBAKHEB, R. (1995). Une comparaison des deux techniques d'estimation des variances: la méthode du jackknife et la méthode du jackknife linéarisée. Direction de la méthodologie, document de travail, DMEM-95-005F. Statistique Canada.
- BINDER, D.A. (1996). Méthodes de linéarisation pour les échantillons à une et deux phases: une approche de type «recette». *Techniques d'enquête*, 22, 17-22.
- CASADY, R.J., et VALLIANT, R. (1993). Propriétés conditionnelles des estimateurs de stratification a posteriori selon la théorie normale. *Techniques d'enquête*, 19, 193-203.
- DEVILLE, J., et SÄRNDAAL, C.E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- HUANG, E.T., et FULLER, W.A. (1978). Nonnegative regression estimation for sample survey data. *Proceedings of the Social Statistics Section, American Statistical Association*, 300-305.
- RAO, J.N.K. (1985). Inférence conditionnelle dans les enquêtes par sondage. *Techniques d'enquête*, 11, 17-35.
- ROYALL, R.M., et CUMBERLAND, W.G. (1981). An empirical study of the ratio estimator and estimator of its variance. *Journal of the American Statistical Association*, 76, 66-88.
- SÄRNDAAL, C.E., SWENSSON, B., et WRETMAN, J. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76, 527-537.
- STATISTIQUE CANADA (1990). *Méthodologie de l'enquête sur la population active du Canada*. N° 71-526 au catalogue.
- VALLIANT, R. (1993). Poststratification and conditional variance estimation. *Journal of the American Statistical Association*, 88, 89-96.
- YUNG, W. (1996). Contributions to poststratification in stratified multi-stage samples. Thèse de doctorat inédite, Carleton University, Ottawa, Canada.



7. CONCLUSIONS

Beebakhee (1995) a appliqué les trois estimateurs de variance  $v_J$ ,  $v_{JL}$  et  $v_L$  à diverses enquêtes-ménages entre-prises par Statistique Canada. Ses résultats empiriques montrent que l'estimateur de variance jackknife linéarisé,  $v_{JL}$ , exige constamment moins de temps et d'argent que l'estimateur de variance jackknife ordinaire  $v_J$  pour les enquêtes, tout en approximant très bien  $v_J$ . Il s'agit de résultats importants sur le plan pratique, car les utilisateurs souhaitent un estimateur plus simple dont les résultats s'approcheront au maximum des résultats de l'estimateur  $v_J$  actuel. L'estimateur de variance linéarisé ordinaire  $v_L$  fonctionne aussi bien que  $v_{JL}$  pour ce qui est du temps et de l'argent investis, mais comparativement à  $v_{JL}$ , les valeurs obtenues ne s'approchent pas autant de celles de  $v_J$ .

Si on s'intéresse surtout à l'estimation des totaux ou des ratios, on pourrait être tenté par l'estimateur de variance jackknife linéarisé  $v_{JL}$ , car il est plus facile à manipuler que l'estimateur de variance jackknife ordinaire  $v_J$  et produit des valeurs proches. Pour les statistiques lisses générales,  $v_{JL}$  a les mêmes inconvénients que l'estimateur de variance linéarisé ordinaire  $v_L$ , car ils exigent tous deux l'établissement d'une formule distincte pour chaque statistique, à l'inverse de  $v_J$ . Pour les propriétés statistiques, la simulation suggère que les trois estimateurs de variance,  $v_J$ ,  $v_{JL}$  et  $v_L$ , se soldent par des résultats analogues. Par contre, l'estimateur jackknife incorrect  $v_J^*$ , qui utilise le même ajustement chaque fois qu'on supprime une grappe, donne de piètres résultats, signe qu'il faut procéder à une nouvelle pondération à chaque suppression.

REMERCIEMENTS

Ce travail n'aurait pu être mené à bien sans une subvention du Conseil de recherches en sciences naturelles et en génie du Canada.

ANNEXE

Preuve que  $v_J(X_j) \approx v_{JL}(X_j)$

Pour obtenir le résultat désiré, il faut d'abord approximer la différence  $A_{-1}^{(g)} - A_{-1}^{-1}$ . Partant de l'identité matricielle

$$(I + P\tilde{Q})^{-1} = I - P(I + \tilde{Q}P)^{-1}\tilde{Q}$$

on obtient

$$A_{-1}^{(g)} - A_{-1}^{-1} = A_{-1}^{-1}[I + (A_{-1}^{(g)} - A_{-1})A_{-1}^{-1}]^{-1} - A_{-1}^{-1} = A_{-1}^{-1}[I - (A_{-1}^{(g)} - A_{-1})]^{-1} - A_{-1}^{-1}$$

$$(I + A_{-1}^{-1}(A_{-1}^{(g)} - A_{-1}))^{-1}A_{-1}^{-1} - A_{-1}^{-1} \approx -A_{-1}^{-1}(A_{-1}^{(g)} - A_{-1})A_{-1}^{-1}. \tag{A.1}$$

Pour cette raison, il est impossible d'obtenir des échantillons mal équilibrés, puisque si  $\hat{M}$  débouche sur une surestimation grossière avec certaines strates, les autres chiffres effectuent la correction à cause de la contrainte décrite plus haut. Par conséquent, on obtient des échantillons passablement bien équilibrés, auquel cas  $v_L$  devrait déboucher sur de bons résultats.

Tableau 7

Biais relatif conditionnel (%) de l'estimateur de régression généralisée

Groupe	$v_L(X_j)$	$v_{JL}(X_j)$	$v_J(X_j)$	$v_J^*(X_j)$
1	9.25	4.95	5.13	26.51
2	3.99	1.50	1.67	24.96
3	-3.24	-4.76	-4.59	17.53
4	-2.66	-3.43	-3.26	20.53
5	7.90	7.61	7.80	35.46
6	-3.60	-3.12	-2.94	23.38
7	-9.24	-8.27	-8.08	17.41
8	3.34	5.30	5.50	35.84
9	-3.75	-0.85	-0.62	30.84
10	-8.68	-4.15	-3.92	28.50

Tableau 8

Taux d'erreur conditionnel (%) de l'estimateur de régression généralisée

Groupe	$v_L(X_j)$	$v_{JL}(X_j)$	$v_J(X_j)$	$v_J^*(X_j)$
1	4.3	4.5	4.4	3.0
2	4.9	5.0	5.0	3.3
3	5.0	5.1	5.1	3.8
4	5.7	5.9	5.9	3.3
5	3.9	4.0	4.0	2.3
6	5.7	5.8	5.7	3.0
7	5.9	5.8	5.8	2.9
8	5.8	5.7	5.7	2.8
9	5.5	5.1	4.9	3.0
10	6.3	5.8	5.8	3.3

Les tableaux 7 et 8 reproduisent les résultats pour l'estimateur de régression généralisée: soit le biais relatif conditionnel au tableau 7 et le taux d'erreur conditionnel (taux nominal de 5%) au tableau 8. Ces résultats ressemblent beaucoup à ceux obtenus de la stratification unidimensionnelle. Dans les deux cas, on remarque une fois de plus la proximité des valeurs de rendement de  $v_J$  et  $v_{JL}$ , ce qui confirme l'équivalence asymptotique de  $v_J$  et  $v_{JL}$ . En résumé, les trois estimateurs de variance  $v_J$ ,  $v_{JL}$  et  $v_L$  produisent à peu près les mêmes résultats. L'estimateur jackknife incorrect  $v_J^*$ , par contre, laisse à désirer, signe qu'il faut effectuer une nouvelle pondération chaque fois qu'on supprime une grappe.

où  $a$  et  $b$  désignent le niveau des deux variables de stratification marginale dans les strates  $a$  et  $b$ . La valeur  $D_p$  donne une idée de la façon dont l'échantillon est «équilibré» par rapport à la répartition de la population marginale dans les strates  $a$  et  $b$ . Les chiffres de population marginaux correspondants. La cation  $a$  et  $b$  désignent le niveau des deux variables de stratification marginale dans les strates  $a$  et  $b$ . La valeur  $D_p$  donne une idée de la façon dont l'échantillon est «équilibré» par rapport à la répartition de la population marginale dans les strates  $a$  et  $b$ . Les chiffres de population marginaux correspondants.

Tableau 5

Biais relatif conditionnel (%) de l'estimateur de stratification a posteriori

Groupe	$v_L (X_{ps})$	$v_{JL} (X_{ps})$	$v_J (X_{ps})$	$v_J^* (X_{ps})$
1	-5.00	-8.05	-7.88	17.83
2	0.55	-1.18	-1.01	28.06
3	8.33	7.03	7.19	41.29
4	-1.10	-1.56	-1.42	31.82
5	-0.76	-0.69	-0.55	34.77
6	2.50	3.39	3.53	41.69
7	6.10	7.51	7.66	48.86
8	6.60	8.82	8.96	53.54
9	-4.46	-1.43	-1.31	41.11
10	-13.56	-9.17	-9.07	36.63

Tableau 6

Taux d'erreur conditionnel (%) de l'estimateur de stratification a posteriori

Groupe	$v_L (X_{ps})$	$v_{JL} (X_{ps})$	$v_J (X_{ps})$	$v_J^* (X_{ps})$
1	5.5	5.9	5.9	3.4
2	4.6	4.8	4.8	2.9
3	3.7	3.8	3.8	1.9
4	5.7	5.8	5.8	2.9
5	4.9	4.8	4.7	2.6
6	5.1	5.0	4.8	2.2
7	5.2	4.8	4.8	2.1
8	4.5	4.3	4.3	1.3
9	5.8	5.4	5.4	2.4
10	7.0	6.3	6.3	2.4

Les tableaux 5 et 6 présentent les résultats pour l'estimateur de stratification a posteriori: soit le biais relatif conditionnel au tableau 5 et le taux d'erreur conditionnel (taux nominal de 5%) au tableau 6. Ces mesures ont été obtenues de la même manière que pour l'essai inconditionnel, mais séparément pour chaque groupe. Les deux tableaux montrent clairement que  $v_J$ ,  $v_{JL}$  et  $v_L$  donnent de bons résultats, bien que ceux de  $v_L$  laissent un peu plus à désirer pour les groupes 1 et 10 aux extrêmes de la fourchette,  $v_J^*$  produit d'assez piètres résultats que dans le cas précédent. Il est quelque peu surprenant de voir  $v_L$  donner de si bons résultats conditionnellement. Une explication plausible est qu'avec ce plan d'échantillonnage particulier, où  $M = \sum (m_{ik})$   $w_{mik} = M$ ,  $\sum^c M = M$ .

Le tableau 4 présente les résultats inconditionnels pour l'estimateur de régression généralisée  $X_p$ . Ainsi qu'on a pu le remarquer avec  $X_{ps}$ , les estimateurs de variance  $v_J$ ,  $v_{JL}$  et  $v_L$  donnent de bons résultats tant en ce qui concerne le biais relatif que le taux d'erreur de l'intervalle de confiance. L'estimateur jackknife incorrect  $v_J^*$ , par contre, débouche sur une sérieuse surestimation, qui se reflète dans un taux d'erreur inférieur au taux nominal et dans un intervalle de confiance de longueur moyenne supérieure.

Tableau 4

Résultats inconditionnels pour l'estimateur de régression généralisée

Mesure de rendement	$v_L (X_p)$	$v_{JL} (X_p)$	$v_J (X_p)$	$v_J^* (X_p)$
Biais relatif (%)	-0.96	0.76	0.57	25.87
Taux d'erreur (%)	5.30	5.27	5.23	3.07
Taux d'erreur minimum (%)	2.24	2.21	2.19	1.08
Taux d'erreur maximum (%)	3.06	3.06	3.04	1.99
Longueur moyenne	3.94	3.95	3.95	4.44

(ii) Résultats conditionnels

Nous nous sommes également attardés aux propriétés conditionnelles des estimateurs de variance, à l'instar de Valliant (1993). Pour l'estimateur de stratification a posteriori, les 10,000 échantillons simulés ont été répartis en 10 groupes comptant chacun 1,000 échantillons avec la mesure suivante (Valliant 1993):

$$D_{ps} = \sum^c \left( \frac{M^c}{M} - 1 \right).$$

On a calculé  $D_{ps}$  pour chaque échantillon et les 10,000 échantillons ont été classés par ordre croissant selon cette valeur, puis divisés en groupes. On peut considérer  $D_{ps}$  comme une indication de l'«équilibre» de l'échantillon par rapport à la distribution de la population dans les strates a posteriori. Pour l'estimateur de régression généralisée, on a retenu l'extension naturelle que voici de  $D_{ps}$ :

$$D_p = \sum^a \left( \frac{M^a}{M} - 1 \right) + \sum^b \left( \frac{M^b}{M} - 1 \right).$$



fois qu'une grappe est supprimée, contrairement à l'estimateur de variance jackknife incorrect. Dans la stratification a posteriori unidimensionnelle,  $v_j^*(Y_{ps}^*)$  applique l'ajustement intégral  ${}^cM/{}^cM$  au lieu de  ${}^cM/{}^cM^{(g)}$ , lorsque la  $(g)$ -ième grappe est supprimée. Bref,  $Y_{ps}^{(g)}$  utilise  $({}^cM/{}^cM)w_{hik}^{(g)}$  comme poids plutôt que  $({}^cM/{}^cM^{(g)})w_{hik}^{(g)}$ . Pareillement, dans la stratification a posteriori bidimensionnelle,  $v_j^*(Y_j^*)$  fait appel à l'ajustement complet  $a_{hik}$  plutôt qu'à  $a_{hik}^{(g)}$ , quand la  $(g)$ -ième grappe disparaît, c'est-à-dire que  $Y_j$  utilise les poids  $w_{hik}^{(g)}$  au lieu de  $w_{hik}^{(g)}$ . La version linéarisée de  $v_j^*$  est identique à l'estimateur de variance  $v_R$  (équation 3.4), dans laquelle  $c_{hik}$  est remplacé par  $y_{hik}$  pour  $X_{ps}^*$ , et à  $v_{JL}$  (équation 4.6) dans laquelle  $e_{hik}$  est remplacé par  $y_{hik}$  pour l'estimateur de régression généralisée  $Y_j$ . En d'autres termes,

$$v_j^*(Y_{ps}^*) = v(y_{hi}^*)$$

où

$$y_{hi}^* = \sum_{c \in c_s} \sum_{k \in k_s} (n_h {}^c w_{hik}) y_{hik}$$

et

$$v_j^*(Y_j^*) = v(y_{hi}^*)$$

où

$$y_{hi}^* = \sum_{k \in k_s} (n_h w_{hik}^*) y_{hik}.$$

Puisque  $v_j^*$  recourt à  $y$  plutôt qu'aux résidus  $e$ ,  $v_j^*$  devrait manifestement surestimer la variance véritable de l'estimateur, même si le calcul s'avère plus simple que pour  $v_j$ .

### (ii) Résultats inconditionnels

Pour comparer l'efficacité inconditionnelle des estimateurs de variance, nous avons calculé le biais relatif empirique (RB) de chaque estimateur de variance. Le RB d'un estimateur de variance  $v$  est donné par:

$$RB = \frac{1}{I} \left[ \frac{MSE}{10,000} \sum_{i=1}^I v_i \right] - 1$$

où  $v_i$  représente la valeur de  $v$  pour le  $i$ -ième échantillon simulé ( $i = 1, \dots, 10,000$ ) et MSE correspond à l'erreur quadratique moyenne (EQM) empirique de l'estimateur, par exemple  $Y_j$ :

$$MSE = \frac{1}{I} \sum_{i=1}^I (Y_i - Y)^2$$

où  $Y_i$  est la valeur de  $Y$  pour le  $i$ -ième échantillon simulé. On a aussi calculé le taux d'erreur des intervalles de confiance théoriques normaux de  $Y$ , dans son ensemble, pour chaque estimateur de variance, au moyen d'un taux d'erreur nominal de 5%:

$$\begin{aligned} \text{taux d'erreur} &= 1 - \frac{1}{10,000} (\text{nombre d'échantillons où } L_i \leq Y \leq U_i), \\ &\text{minimum et maximum sont:} \\ \text{taux d'erreur minimum} &= \frac{1}{10,000} (\text{nombre d'échantillons où } Y < L_i) \\ \text{taux d'erreur maximum} &= \frac{1}{10,000} (\text{nombre d'échantillons où } Y > U_i). \end{aligned}$$

La longueur moyenne de l'intervalle de confiance correspond à:

$$\text{longueur moyenne} = \frac{1}{I} \sum_{i=1}^I (U_i - L_i).$$

Le tableau 3 reproduit des résultats inconditionnels pour l'estimateur de stratification a posteriori  $X_{ps}^*$  obtenus au moyen des mesures de rendement qui précèdent. En ce qui concerne le biais relatif,  $v_{JL}$  et  $v_j$  donnent de bons résultats quand  $RB < 1\%$ , mais l'estimateur de variance jackknife incorrect  $v_{JL}^*$  surestime de façon appréciable l'EQM ( $RB = 37\%$ ). Notons que  $v_{JL}$  donne aussi une bonne estimation de l'EQM de  $X_{ps}^*$  de façon inconditionnelle ( $RB < 1\%$ ), contrairement à ce que prétend Valliant (1993). Ce dernier avait signalé un biais relatif de 35% pour  $v_{JL}$  avec les mêmes données. Face à la convergence de  $v_{JL}$  selon le plan d'échantillonnage, qu'appuient les résultats de la simulation pour  $v_{JL}$ , on peut supposer que les calculs de Valliant pour  $v_{JL}$  sont erronés.

Tableau 3

Résultats inconditionnels pour l'estimateur de stratification a posteriori

Mesure de rendement	$v_{JL}(X_{ps}^*)$	$v_j(X_{ps}^*)$	$v_j(X_{ps}^*)$	$v_j^*(X_{ps}^*)$
Biais relatif (%)	-0.44	0.12	5.09	0.26
Taux d'erreur (%)	5.20	2.35	5.06	37.16
Taux d'erreur minimum (%)	2.41	2.33	2.33	0.99
Taux d'erreur maximum (%)	2.79	2.74	2.73	1.42
Longueur moyenne	3.81	3.82	3.83	4.48

En ce qui concerne le rendement relatif à l'intervalle de confiance, le tableau 3 montre que le taux d'erreur associé à  $v_j$ , à  $v_{JL}$  et à  $v_{JL}^*$  s'approche du taux nominal de 5%, tandis que le taux d'erreur de  $v_j^*$  est considérablement plus faible (environ 2.5%). On remarque une performance

$$Y_i = Y + (X - X) ^T B_1$$

un estimateur de régression généralisée de la somme totale allouée à l'achat de vêtements,  $Y$ . De même, soit

$$Z_i = Z + (X - X) ^T B_2$$

un estimateur de régression généralisée du revenu total,  $Z$ . La proportion à laquelle on s'intéresse est exprimée par  $\theta = Y/Z$ , qu'on peut estimer par

$$\hat{\theta} = Y_i / Z_i.$$

L'estimateur de variance jackknife est représenté par

$$v_j(\theta) = \sum_g \frac{n_g}{n_g - 1} \sum_j (\hat{\theta}_{(g)} - \theta)^2 \tag{5.1}$$

où

$$\hat{\theta}_{(g)} = Y_{r(g)} / Z_{r(g)}.$$

Lorsqu'on linéarise l'estimateur (5.1), on en obtient un nouveau

$$v_{jL}(\theta) = v(r_{hi}^{**}) \tag{5.2}$$

où

$$r_{hi}^{**} = \frac{1}{L} \sum_k (n_h w_{hik}^*) e_{hik}^*$$

pour laquelle

$$e_{hik}^* = e_{hik} - \frac{Z_r}{Y_r} e_{hik},$$

et

$$e_{hik} = y_{hik} - x_{hik}^T B_1, \quad \hat{e}_{hik} = z_{hik} - x_{hik}^T B_2.$$

La preuve de (5.2) est omise pour plus de simplicité.

6. SIMULATION

On a procédé à une simulation afin de vérifier les propriétés conditionnelles et inconditionnelles des estimateurs de variance pour un échantillon de population finie avec une et deux variables de stratification a posteriori. On a recouru pour cela à une population fixe finie, déjà envisagée par Valliant (1993), soit 10,841 personnes de l'Enquête sur la population courante (EPC) américaine de septembre 1988. La variable qui nous intéresse,  $y$ , correspond à la rémunération hebdomadaire de chaque sujet. La stratification a posteriori à une dimension reposait sur l'âge, la race et le sexe, alors que celle à deux dimensions portait sur cinq niveaux d'âge et deux niveaux de race (voir tableaux 1 et 2 pour plus de détails).

Note: Le numéro en marge correspond au numéro d'identification des strates. Les cellules ( $ij$ ) indiquent la strate ( $i = 1, \dots, 5; j = 1, 2$ ).

Age	Autre race		Race noire	
	PS2(1)	PS2(2)	PS1(1)	PS1(2)
19 ans et moins	(1,1)	(1,2)	PS1(1)	PS1(2)
20-24	(2,1)	(2,2)	PS1(2)	PS1(3)
25-34	(3,1)	(3,2)	PS1(3)	PS1(4)
35-64	(4,1)	(4,2)	PS1(4)	PS1(5)
65 ans et plus	(5,1)	(5,2)	PS1(5)	

Tableau 2

Attribution des catégories âge/race aux strates a posteriori: stratification a posteriori bidimensionnelle

Note: Le numéro des cellules (1-8) correspond au numéro d'identification des strates.

Age	Autre race		Race noire	
	Homme	Femme	Homme	Femme
19 ans et moins	1	1	1	1
20-24	2	3	3	3
25-34	5	6	4	4
35-64	7	8	4	4
65 ans et plus	2	3	3	1

Tableau 1

Attribution des catégories âge/race/sexe aux strates a posteriori: stratification a posteriori unidimensionnelle

Nous avons calculé l'estimateur de base, l'estimateur de stratification a posteriori pertinent,  $\hat{Y}_{ps}$  ou  $\hat{Y}_r$ , et quatre estimateurs de variance pour chaque échantillon: l'estimateur de variance linéarisé ordinaire  $v_L$ , l'estimateur de variance jackknife linéarisé  $v_{jL}$ , l'estimateur de variance jackknife  $v_j$  et un estimateur de variance jackknife incorrect  $v_j^*$ . Quand on applique la méthode du jackknife, une question se pose: faut-il ou non calculer à nouveau les poids «finals» ou «étalonnés» chaque fois qu'une grappe est supprimée? Le bon estimateur de variance jackknife donne effectivement un nouveau poids «final» chaque



Quand on passe à l'estimation de la variance, l'équation (2.2) donne une fois de plus l'estimateur de variance linéaire ordinaire où  $y_{hi}$  devient

$$\bar{e}_{hi} = \sum_k (n_h w_{hik}^*) e_{hik},$$

où

$$e_{hik} = y_{hik} - \mathbf{x}_{hik}^T \mathbf{B} \quad (4.3)$$

correspond aux résidus estimatifs, c'est-à-dire

$$v_L(Y_j) = v(\bar{e}_{hj}). \quad (4.4)$$

Pour la méthode du jackknife, il est nécessaire de calculer les poids d'échantillonnage  $w_{hik}^*$  chaque fois qu'on supprime une grappe ( $g_j$ ). Ces poids sont donnés par

$$w_{hik}^*(g_j) = w_{hik}(g_j) a_{hik}(g_j),$$

où

$$a_{hik}(g_j) = 1 + \mathbf{x}_{hik}^T \mathbf{A}_{-1}^{-1}(g_j) (\mathbf{X} - \mathbf{X}^{(g_j)}),$$

$$\mathbf{A}(g_j) = \sum_{(hik) \in s} w_{hik}(g_j) \mathbf{x}_{hik} \mathbf{x}_{hik}^T,$$

et

$$\mathbf{X}(g_j) = \sum_{(hik) \in s} w_{hik}(g_j) \mathbf{x}_{hik}.$$

L'estimateur de régression généralisée résultant se présente de la façon suivante:

$$Y_j^{(g_j)} = \sum_{(hik) \in s} w_{hik}^*(g_j) y_{hik}$$

$$= Y_j^{(g_j)} + (\mathbf{X} - \mathbf{X}^{(g_j)})^T \mathbf{B}(g_j)$$

où  $\mathbf{B}(g_j)$  est le vecteur des coefficients de régression estimatifs lorsque la ( $g_j$ )-ième grappe est supprimée:

$$\mathbf{B}(g_j) = \mathbf{A}_{-1}^{-1}(g_j) \mathbf{b}(g_j)$$

avec

$$\mathbf{b}(g_j) = \sum_{(hik) \in s} w_{hik}(g_j) \mathbf{x}_{hik} y_{hik}.$$

L'estimateur de variance jackknife de  $Y_j$  est donc représenté par

$$v_j(Y_j) = \sum_{g=1}^L \frac{n_g}{n_g - 1} \sum_{j=1}^r (Y_j^{(g_j)} - Y_j)^2. \quad (4.5)$$

On verra à l'annexe qu'en linéarisant l'estimateur de variance jackknife (4.5), on obtient

$$e_{hi}^* = \sum_k (n_h w_{hik}^*) e_{hik}$$

pour laquelle

$$v_L(Y_j) = v(e_{hi}^*). \quad (4.6)$$

où  $w_{hik}^*$  est défini en (4.2) et  $e_{hik}$ , en (4.3). Fait intéressant, l'estimateur de variance jackknife linéarisé (4.6) ressemble à l'estimateur de variance fondé sur un modèle que proposent Särndal, Swensson et Wretman (1989) dans le cadre d'un échantillonnage à un degré. Yung (1996) a établi l'équivalence asymptotique de  $v_j(Y_j)$  et  $v_{jL}(Y_j)$  aux ordres supérieurs dans le cas particulier mais important où on retrouve  $n_h = 2$  grappes par strate. Souignons que les résultats qui précèdent s'appliquent aussi aux variables auxiliaires générales  $\mathbf{x}_{hik}$ .

Binder (1996) propose une nouvelle méthode de linéarisation qui aboutit elle aussi à  $v_{jL}(Y_j)$ . En vertu de cette méthode, on évalue les dérivées partielles par rapport aux valeurs estimatives  $Y_j$ ,  $\mathbf{X}$  et  $\mathbf{B}$ , et non par rapport aux chiffres de population  $Y_j$ ,  $\mathbf{X}$  et  $\mathbf{B}$  comme avec la méthode de linéarisation classique. Puisque  $v_j$  et  $v_{jL}$  sont convergents selon le plan d'échantillonnage (Yung 1996) et présentent de bonnes propriétés conditionnelles, nos résultats apportent une justification théorique à la méthode de type «recette» que propose Binder.

Le calcul de l'estimateur de variance jackknife suppose l'inversion de la matrice  $\mathbf{A}(g_j)$  pour chaque ( $g_j$ ). Toute-

fois, on peut obtenir la valeur approximative de cet estimateur en gardant l'inverse pour l'échantillon complet,  $\mathbf{A}^{-1}$ , puis en utilisant les poids modifiés

$$\tilde{w}_{hik}(g_j) = w_{hik}(g_j) \tilde{a}_{hik}(g_j)$$

où

$$\tilde{a}_{hik}(g_j) = 1 + (w_{hik}/w_{hik}(g_j)) \mathbf{x}_{hik}^T \mathbf{A}_{-1}^{-1} (\mathbf{X} - \mathbf{X}^{(g_j)}).$$

L'estimateur résultant de  $Y_j$  à la suppression de la ( $g_j$ )-ième grappe, est donné par

$$Y_j^{(g_j)} = \sum_{(hik) \in s} \tilde{w}_{hik}(g_j) y_{hik}$$

et l'estimateur de variance jackknife correspondant est

$$v_{j1}(Y_j) = \sum_{g=1}^L \frac{n_g}{n_g - 1} \sum_{j=1}^r (Y_j^{(g_j)} - Y_j)^2. \quad (4.7)$$

Il est facile de voir que (4.7) est identique à l'estimateur de variance linéarisé ordinaire (4.4).

## 5. ESTIMATION D'UN RATIO

On a souvent besoin d'établir le ratio de deux totaux estimatifs. Dans le cadre d'une enquête sur les dépenses des familles, par exemple, on s'intéresse à la proportion du revenu consacré à l'habillement. Soit:

où

$$e_{hi}^* = \sum_{k \in c_s} \sum_{h \in c_{whik}} (n_h {}^c w_{hik}) \cdot e_{hik}.$$

Avec la méthode du jackknife, il est nécessaire de recalculer les poids de stratification a posteriori  ${}^c w_{hik}$  chaque fois qu'on supprime une grappe ( $g_j$ ). On se sert donc des poids jackknife  $w_{hik(g_j)}$  de (3.1) pour obtenir  ${}^c M_{(g_j)}$  on utilise  ${}^c w_{hik(g_j)} = ({}^c M / {}^c M_{(g_j)}) w_{hik(g_j)}$  afin d'établir

$$Y_{ps(g_j)} = \sum_{h \in c_s} \sum_{h \in c_{whik(g_j)}} {}^c w_{hik(g_j)} Y_{hik}.$$

L'estimateur de variance jackknife devient donc:

$$v_j(Y_{ps}) = \sum_{g=1}^G \frac{n_g}{n_g - 1} \sum_{h=1}^{n_g} (Y_{ps(g_j)} - Y_{ps})^2. \quad (3.5)$$

En linéarisant (3.5), on parvient à un estimateur de variance jackknife linéarisé,  $v_{jL}(Y_{ps})$ , identique à l'estimateur de variance de Rao (3.4); lire aussi Valliant (1993). Dans le cas particulier, non moins important, où  $n_h = 2$  grappes par strate, les équations (3.4) et (3.5) sont asymptotiquement égales en tenant compte de termes d'ordre supérieur, lorsque le nombre de strates  $L$  augmente (Yung 1996).

Rao (1985) justifie (3.4) par des arguments heuristiques en soulignant que l'équation se réduit à un estimateur de variance valable pour un échantillonnage aléatoire simple, étant donné les tailles d'échantillon dans les strates a posteriori, à l'inverse de l'estimateur de variance linéarisé ordinaire (3.3). Särndal, Swensson et Wretman (1994) sont parvenus à un estimateur de variance similaire à (3.4) avec un échantillonnage à un degré dont le cadre était articulé sur un modèle. Puisque  $v_{jL}(Y_{ps})$  et  $v_j(Y_{ps})$  sont presque égaux, les résultats qui précèdent suggèrent deux estimateurs de variance «robustes» selon la taille estimée des strates a posteriori. Valliant (1993) a effectué une simulation pour éprouver la «robustesse» de  $v_j(Y_{ps})$  et  $v_{jL}(Y_{ps})$ .

#### 4. ESTIMATEUR DE RÉGRESSION GÉNÉRALISÉE

Dans la pratique, on crée couramment des strates a posteriori en fonction de deux variables auxiliaires ou d'avantage. Sachant le chiffre de population résultant pour chaque cellule, on peut se servir de l'estimateur de stratification a posteriori ajusté par ratio pour rendre les estimations plus précises. En réalité cependant, il se pourrait que le chiffre de population de la cellule soit inconnu. On pourrait, par exemple savoir le chiffre marginal de certains groupes d'âge et groupes ethniques, mais pas le chiffre des cellules combinant âge et race. Dans un tableau bidimensionnel, il s'ensuivrait donc qu'on connaîtrait les chiffres marginaux, mais pas les chiffres par cellule. Lorsque les paramètres de stratification a posteriori sont

plus nombreux et qu'on connaît les chiffres de population marginaux, on peut se servir d'un estimateur de régression généralisée de  $Y$  en utilisant des variables auxiliaires indicatrices pour désigner les catégories des paramètres de stratification a posteriori (Huang et Fuller 1978; Deville et Särndal 1992).

Soit  $x_{hik}$ , un vecteur des variables auxiliaires dont la population totale connue est  $X$ . L'estimateur de régression généralisée de  $Y$  s'exprime comme suit

$$Y_i = Y + (X - X)^T B, \quad (4.1)$$

où

$$X = \sum_{h \in c_s} w_{hik} x_{hik},$$

et où  $B$  représente le vecteur des coefficients de régression estimatifs

$$B = A^{-1} b,$$

où

$$A = \sum_{h \in c_s} w_{hik} x_{hik} x_{hik}^T,$$

et

$$b = \sum_{h \in c_s} w_{hik} x_{hik} Y_{hik}.$$

L'estimateur de stratification a posteriori  $Y_{ps}$  constitue un cas particulier de (4.1) où  $x_{hik}$  indique le vecteur des variables indicatrices des strates a posteriori. Dans ce cas,  $X = ({}^1 M, \dots, {}^c M)^T$ ,  $X = ({}^1 M, \dots, {}^c M)^T$  et  $B = ({}^1 R, \dots, {}^c R)^T$ , où  ${}^c R = {}^c Y / {}^c M$ . Par conséquent,

$$Y_i = Y + \sum_{c \in c_s} {}^c R ({}^c M - {}^c M) = Y_{ps}.$$

Avec deux paramètres de stratification a posteriori ou plus,  $X$  correspond à la valeur vectorielle des données de population marginales. On peut récrire l'estimateur de régression généralisée de la façon suivante:

$$Y_i = \sum_{h \in c_s} w_{hik}^* Y_{hik},$$

où

$$w_{hik}^* = w_{hik} a_{hik}$$

correspond au poids «final» ou au poids «d'étalonnage» et

$$a_{hik} = 1 + x_{hik}^T A^{-1} (X - X).$$

Dans le cas particulier  $Y_{ps}$ ,  $a_{hik} = {}^c M / {}^c M$  pour  $(hik) \in c_s$ . En remplaçant  $Y_i$  dans l'opérateur par  $Y_i(Y_{hik})$ , on se généralisée  $X_i = Y_i(x_{hik}) = X$ , ce qui garantit la cohérence avec les totaux connus  $X$ .



## 2. ESTIMATEUR DE BASE

Partant des poids de base  $w_{hik}$ , l'estimateur non biaisé de la population totale  $Y$  s'écrit

$$\bar{Y} = \sum_{(hik) \in s} w_{hik} y_{hik}, \quad (2.1)$$

où  $s$  représente l'échantillon d'éléments et  $y_{hik}$  correspond à la valeur de la caractéristique à laquelle on s'intéresse, associée à l'élément  $(hik) \in s$  de l'échantillon. Pour plus de simplicité, on suppose qu'il n'y a que des répondants. La pratique courante consiste à prélever des grappes sans remise. Néanmoins, lorsqu'on doit estimer la variance, on simplifie considérablement les calculs en traitant l'échantillon comme si les grappes avaient été choisies avec remise. Pareille approximation entraîne habituellement une surestimation de la variance de  $\bar{Y}$ , mais le biais relatif sera probablement faible si les taux d'échantillonnage au premier degré le sont également.

L'estimateur de la variance de  $\bar{Y}$  s'exprime sous la forme suivante:

$$v(\bar{Y}) = \sum_L \frac{n_h(n_h - 1)}{1} \sum_{i=1}^{h-1} (y_{hi} - \bar{y}_h)^2 = v(y_{hi}), \quad (2.2)$$

où  $y_{hi} = \sum_k (n_h w_{hik}) y_{hik}$  et  $\bar{y}_h = (1/n_h) \sum_i y_{hi}$ . La notation  $v(y_{hi})$  de l'opérateur indique que  $v(\bar{Y})$  ne dépend que de  $y_{hi}$ .  
Pour introduire la méthode du jackknife, il faut utiliser l'estimateur  $\bar{Y}^{(gj)}$  pour chaque  $(gj)$  de l'échantillon après avoir omis les données de la  $j$ -ième grappe prélevée dans la  $g$ -ième strate ( $j = 1, \dots, n_g; g = 1, \dots, L$ ). Partant de (2.1), on y parvient simplement en supposant que  $w_{gik} = 0$ , en remplaçant  $w_{gik}$  ( $i \neq j$ ) par  $n_g w_{gik} / (n_g - 1)$  et en gardant les poids originaux  $w_{hik}$  pour  $h \neq g$ , c'est-à-dire:

$$w_{hik(g)} = \begin{cases} 0 & \text{si } (hi) = (gj) \\ \frac{n_g}{n_g - 1} w_{gik} & \text{si } h = g \text{ et } i \neq j \\ w_{hik} & \text{si } h \neq g. \end{cases}$$

Ces poids jackknife  $w_{hik(g)}$  sont établis pour chaque grappe  $(gj)$ . L'estimateur résultant de  $\bar{Y}$  est

$$\bar{Y}^{(gj)} = \sum_{(hik) \in s} w_{hik(g)} y_{hik}.$$

L'estimateur de variance jackknife prend la forme suivante:

$$v_J(\bar{Y}) = \sum_L \frac{n_g}{n_g - 1} \sum_{j=1}^{n_g} (\bar{Y}^{(gj)} - \bar{Y})^2. \quad (2.3)$$

## 3. ESTIMATEUR DE STRATIFICATION A POSTERIORI

Supposons que la population soit répartie a posteriori en  $C$  strates dont la population connue est  ${}^cM$ ,  $c = 1, \dots, C$ . Nous nous servirons de l'indice antérieur  $c$  pour désigner les strates a posteriori. Un estimateur de  ${}^cM$  est donné par:

$${}^cM = \sum_{(hik) \in {}^c s} w_{hik}, \quad (3.1)$$

où  ${}^c s$  représente l'échantillon d'éléments appartenant à la  $c$ -ième strate a posteriori. De même, l'estimateur du total de la strate a posteriori  ${}^c Y$  est

$${}^c Y = \sum_{(hik) \in {}^c s} w_{hik} y_{hik}.$$

En utilisant les estimateurs  ${}^c Y$  et  ${}^c M$ , on obtient l'estimateur de stratification a posteriori du total  $Y$ :

$$Y_{ps} = \sum_c \frac{{}^c M}{{}^c Y} {}^c Y. \quad (3.2)$$

L'équation qui précède peut être réécrite de la façon suivante:

$$Y_{ps} = \sum_c \sum_{(hik) \in {}^c s} {}^c w_{hik} y_{hik}$$

où  ${}^c w_{hik} = w_{hik} ({}^c M / {}^c Y)$  correspond au poids de  $(hik) \in {}^c s$  ajusté par ratio. Si  $y_{hik}$  représente la variable indicatrice d'une strate a posteriori, par exemple  $c$ , alors  $Y_{ps} = {}^c M$ , ce qui garantit la cohérence avec les totaux connus,  ${}^c M$ . L'équation (2.2) donne l'estimateur de variance linéarisé ordinaire quand on remplace  $y_{hi}$  par

$$\tilde{e}_{hi} = \sum_c \sum_{k \in {}^c s} (n_h w_{hik}) {}^c e_{hik},$$

où  ${}^c e_{hik} = y_{hik} - {}^c Y / {}^c M$  pour le  $k$ -ième élément de la  $(hi)$ -ième grappe de  ${}^c s$ , c'est-à-dire:

$$v_L(Y_{ps}) = v(\tilde{e}_{hi}). \quad (3.3)$$

Rao (1985) propose un autre estimateur de variance linéarisé reposant sur les poids  ${}^c w_{hik}$  ajustés par ratio:

$$v_R(Y_{ps}) = v(e_{hi}^*). \quad (3.4)$$

# Linéarisation des estimateurs de variance jackknife dans un échantillonnage stratifié à degrés multiples

W. YUNG et J.N.K. RAO<sup>1</sup>

## RÉSUMÉ

Les auteurs examinent l'estimation de la variance d'une totalisation issue d'un échantillonnage stratifié à degrés multiples pour l'estimateur de stratification a posteriori et l'estimateur de régression généralisée. En linéarisant l'estimateur de variance jackknife, on obtient un nouvel estimateur, différent de celui obtenu par la méthode de linéarisation ordinaire. En matière de calcul, cet estimateur est plus simple à utiliser que l'estimateur de variance jackknife. Pourtant, il donne des valeurs qui s'approchent de celles de la méthode du jackknife. Les auteurs étudient les propriétés de l'estimateur de variance jackknife linéarisé, de l'estimateur de variance linéarisé ordinaire et de l'estimateur de variance jackknife dans le cadre d'une simulation. D'après l'écart entre le total estimatif des variables auxiliaires et les totaux connus de la population, les trois estimateurs donnent de bons résultats, conditionnellement ou non. Un estimateur de variance jackknife reposant sur une nouvelle pondération incorrecte a donné de piètres résultats, signe qu'il est important de procéder de façon adéquate à une nouvelle pondération quand on recourt à la méthode du jackknife.

MOTS CLÉS: Estimateur de régression généralisée; estimateur de variance jackknife; estimateur de variance linéarisé; estimateur de stratification a posteriori.

## 1. INTRODUCTION

Les enquêtes-échantillons de grande envergure recourent souvent à des plans d'échantillonnage stratifiés à plusieurs degrés et un nombre important de strates,  $L$ , ainsi que relativement peu d'unités primaires d'échantillonnage (grappes),  $n_h$  ( $\geq 2$ ), dans chaque strate. On prélève certains éléments (unités finales d'échantillonnage) dans chaque grappe au moyen d'une méthode d'échantillonnage. Le nombre de degrés ou les méthodes d'échantillonnage utilisées après le premier degré de l'échantillonnage ne sont pas spécifiques, mais on suppose que le sous-échantillonnage à l'intérieur des grappes débouche sur une estimation non biaisée des valeurs totales des grappes,  $Y_{hi}$ ,  $i = 1, \dots, n_h$ .

Les spécifications du plan d'échantillonnage de l'enquête servent à établir les poids de base  $w_{hi}$  ( $> 0$ ) associés au ( $hik$ )-ième élément. Les poids de base  $w_{hi}$  doivent souvent être ajustés après stratification pour être cohérents avec les totaux connus des variables de stratification a posteriori. En présence d'un seul paramètre de stratification a posteriori, on ajuste les poids par ratio à la population connue (à savoir, chiffres selon l'âge et le sexe). Avec deux paramètres de stratification a posteriori ou plus et des chiffres de population marginaux connus, on peut étalonner les poids  $w_{hi}$  par régression généralisée (voir partie 4), comme on le fait dans l'Enquête sur la population active du Canada (EPA). L'EPA recourt à la méthode du jackknife pour estimer la variance de l'estimateur de régression généralisée. Cette méthode exige des calculs importants à l'ordinateur, mais elle s'applique aisément aux statistiques généralement

lisses, contrairement à la linéarisation. Par ailleurs, elle présente de bonnes propriétés conditionnelles. Par exemple, avec l'échantillonnage aléatoire simple et l'estimateur de ratio, Royall et Cumberland (1981) ont montré que l'estimateur de variance jackknife suit la variance conditionnelle, étant donné la moyenne de la variable auxiliaire  $x$  pour l'échantillon. La présente communication a principalement pour but d'examiner l'estimation de la variance pour l'estimateur de stratification a posteriori ajusté par quotient et pour l'estimateur de régression généralisée, avec un échantillonnage stratifié. En linéarisant l'estimateur de variance jackknife, on obtient un nouvel estimateur, différent de l'estimateur de variance linéarisé ordinaire. Dans le cas de l'estimateur de stratification a posteriori, ce nouvel estimateur est identique à l'estimateur de variance de Rao (1985). Le nouvel estimateur de variance proposé est plus simple à calculer que l'estimateur de variance jackknife, mais donne des résultats similaires. La partie 2 présente l'estimateur de variance jackknife pour l'estimateur de base d'un total,  $Y$ . Les estimateurs de variance jackknife et jackknife linéarisé pour l'estimateur de stratification a posteriori sont présentés à la partie 3. Les résultats sont étendus à l'estimateur de régression généralisée pour les variables de stratification multiple a posteriori à la partie 4. La partie 5 traite de l'estimation de la variance pour un ratio de deux totaux, obtenu chacun au moyen de l'estimateur de régression généralisée. Les résultats d'une simulation démontrant l'efficacité relative de l'estimateur de variance linéarisé ordinaire, de la méthode du jackknife et de l'estimateur de variance jackknife linéarisé apparaissent à la partie 6.

<sup>1</sup> W. Yung, Statistique Canada, Division des méthodes d'enquêtes-ménage, immeuble R.H. Coats, parc Tunney, Ottawa, Ontario, K1A 0T6; et J.N.K. Rao, Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, K1S 5B6.



$$U_1(\theta, \hat{\lambda}_\theta) = \sum_{k \in s} w_k u_1(y_k, \theta, \hat{\lambda}_\theta) = 0,$$

où  $\hat{\lambda}_\theta$  est l'estimation d'un paramètre additionnel, défini comme étant la solution à:

$$U_2(\theta, \hat{\lambda}_\theta) = \sum_{k \in s} w_k u_2(y_k, \theta, \hat{\lambda}_\theta) = 0,$$

pour une valeur  $\theta$  donnée. Par un argument basé sur l'élimination des paramètres additionnels pour les problèmes de test d'hypothèse sur  $\theta$ , Binder et Patak recommandent de baser les inférences au sujet de  $\theta$  sur la variable

$$u^* = u_1(y, \theta, \hat{\lambda}_\theta) - \left[ \frac{\partial U_1}{\partial \hat{\lambda}_\theta} \right] \left[ \frac{\partial U_2}{\partial \hat{\lambda}_\theta} \right]^{-1} u_2(y, \theta, \hat{\lambda}_\theta). \quad (17)$$

En particulier, les intervalles de confiance bilatéraux pour  $\theta$  devraient être basés sur:

$$\left\{ \theta \left| \frac{W}{U_1^2(\theta, \hat{\lambda}_\theta)} \leq \chi_{2-\alpha}^2(1) \right. \right\},$$

où  $W$  est la variance estimée de l'estimateur d'un total lorsque la variable estimée est  $u^*$ .

Posons  $u_1 = g(\lambda_1, \lambda_2) - \theta$ . Le noyau des équations d'estimation pour les totaux sur  $y$  sera donné par  $u_{21} = y - \lambda_1$  et le noyau des équations d'estimation pour  $u_{22}(\lambda_1, \lambda_2)$ . Posons

$$U_2 = \sum w_k \begin{bmatrix} u_{21} \\ u_{22} \end{bmatrix} = \begin{bmatrix} Y - N\hat{\lambda}_1 \\ Nu_{22} \end{bmatrix}, \quad \text{où } N = \sum w_k.$$

Après quelques manipulations algébriques, nous obtenons, de (17), la variance qui nous intéresse, soit la variance du total estimé basé sur la variable  $u^*$ , donnée par:

## BIBLIOGRAPHIE

- BINDER, D.A., et PATAK, Z. (1994). Use of estimating functions for interval estimation from complex surveys. *Journal of the American Statistical Association*, 89, 1035-1043.
- KOTT, P.S. (1990). Estimating the conditional variance of a design consistent regression estimator. *Journal of Statistical Planning and Inference*, 24, 287-296.
- RAO, J.N.K. (1995). Communication privée.
- RAO, J.N.K., et SITTER, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82, 453-460.
- SÄRNDAAL, C.-E., SWENSSON, B., et WRETMAN, J.H. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76, 527-537.

## REMERCIEMENTS

Je désire remercier Georgia Roberts et Alain Théberge, dont les suggestions nombreuses et utiles ont permis d'améliorer la lisibilité d'une ébauche antérieure. Je remercie également J.N.K. Rao pour les discussions nombreuses et utiles sur ce sujet, ainsi qu'un arbitre anonyme pour ses remarques constructives.

Ceci est équivalent à l'expression (10), ce qui démontre que les méthodes démontrées ici sont cohérentes avec celles de Binder et Patak (1994).

+ des termes constant.

$$\left[ \frac{\partial g(\hat{\lambda}_1, \hat{\lambda}_2)}{\partial \hat{\lambda}_1} \right]'_y - \left[ \frac{\partial g(\hat{\lambda}_1, \hat{\lambda}_2)}{\partial \hat{\lambda}_2} \right]' \left[ \frac{\partial u_{22}(\hat{\lambda}_1, \hat{\lambda}_2)}{\partial \hat{\lambda}_2} \right]^{-1} \left[ \frac{\partial u_{22}(\hat{\lambda}_1, \hat{\lambda}_2)}{\partial \hat{\lambda}_1} \right]'_y$$

### 3. ÉCHANTILLONS À DEUX PHASES

La méthode décrite ci-dessus s'étend assez facilement au cas des échantillons à deux phases. Par exemple, examinons l'estimateur par quotient à deux phases d'une population totale, donné par:

$$Y_{R(2)} = \frac{X}{Y} X_{(1)} = R X_{(1)}, \quad (15)$$

où  $X_{(1)} = \sum w_k x_k$  est l'estimation de la première phase de  $X$  basée sur les poids de la première phase  $\{w_k\}$ , et  $Y$  sont les estimations de  $Y$  et  $X$ , respectivement, basées sur les unités d'échantillonnage de la deuxième phase avec les poids  $\{w_k w_{2k}\}$ , où  $w_{2k}$  est le poids assigné à l'unité de la deuxième phase choisie et est soumise à la condition de se trouver dans l'échantillon de la première phase. En particulier, posons:

$$a_k = \begin{cases} 1 & \text{si la } k\text{-ième unité est dans l'échantillon de la} \\ & \text{seconde phase} \\ 0 & \text{autrement,} \end{cases}$$

nous obtenons,

$$Y = \sum_{k \in s} w_k w_{2k} a_k y_k,$$

où  $s$  est l'ensemble des indices correspondant aux unités de l'échantillon de la première phase. En prenant les différentielles totales de (15), nous obtenons:

$$(dY_{R(2)}) = \left( \frac{X}{X_{(1)}} \right) [(dY) - R(dX)] + R(dX_{(1)}).$$

Nous remplaçons maintenant les différentielles totales par les sommes pondérées sur les unités de la première phase:

$$Y_{R(2)} - Y =$$

$$\sum_{k \in s} w_k \left[ a_k w_{2k} \left( \frac{X}{X_{(1)}} \right) (y_k - R x_k) + R x_k \right] + \dots,$$

de sorte que

$$z_k = a_k w_{2k} \left( \frac{X}{X_{(1)}} \right) (y_k - R x_k) + R x_k. \quad (16)$$

Nous voyons que les étapes que nous avons suivies sont essentiellement les mêmes que dans le cas d'un échantillon à une phase. Toutefois, il importe de souligner que  $z_k$  contient maintenant la variable aléatoire  $a_k$ , que l'on utilise pour indiquer si l'unité échantillonnée se trouve ou non dans l'échantillon de la deuxième phase. Cette variable est nécessaire pour calculer l'estimateur de la variance à deux phases.

### 4. JUSTIFICATION

Il est alors nécessaire de placer la variable  $z$  dans l'algorithme qui calcule la variance de l'estimateur d'un total provenant d'un échantillon à deux phases.

$$z_k = \left[ \frac{\partial g}{\partial X_{(1)}} \right]' x_k - \left[ \frac{\partial g}{\partial \lambda} \right]' \left[ \frac{\partial \lambda}{\partial U} \right]^{-1} \left[ \frac{\partial \lambda}{\partial Y} \right] a_k w_{2k} y_k.$$

Ainsi, l'expression générale pour  $z_k$  est:

$$T - T = \left[ \frac{\partial g}{\partial X_{(1)}} \right]' \left( \sum_{k \in s} w_k x_k - X \right) - \left[ \frac{\partial g}{\partial \lambda} \right]' \left[ \frac{\partial \lambda}{\partial U} \right]^{-1} \left[ \frac{\partial \lambda}{\partial Y} \right] \left( \sum_{k \in s} a_k w_k w_{2k} y_k - Y \right).$$

de sorte que:

$$(d\lambda) = - \left[ \frac{\partial \lambda}{\partial U} \right]^{-1} \left[ \frac{\partial \lambda}{\partial Y} \right] (dY),$$

En prenant les différentielles totales, nous obtenons, comme dans l'expression (9):

$$T = g(X_{(1)}, \lambda).$$

et

$$U(\lambda, Y) = 0,$$

L'extension de cette méthode aux autres problèmes d'estimation dans les échantillons à deux phases ne présente pas de problème. Supposons, par exemple, que  $(Y_1, \dots, Y_m)$  soient des estimations de  $(X_1, \dots, X_m)$  pour les échantillons de la deuxième phase, et que  $(X_{(1)}^1, \dots, X_{(1)}^m)$  soient des estimations des variables disponibles seulement pour les unités de l'échantillon de la première phase. Nous supposons qu'un ensemble de paramètres additionnels,  $\lambda$ , soit défini seulement en termes des unités dans la deuxième phase, et que la variable qui nous intéresse soit définie en termes de ces paramètres additionnels et des valeurs  $X_{(1)}^j$ . Nous obtenons donc l'expression formelle suivante:

La technique que nous venons de décrire peut être considérée comme un résultat direct de la formulation donnée par Binder et Patak (1994). Nous résumerons ici l'un des principaux résultats de cette communication. Supposons que nous soyons intéressés par le paramètre  $\theta$ , défini comme étant la solution à:



$$(dY^{GREG}) = (dY) - \hat{\beta}'(dX) + (d\hat{\beta})'(X - X)$$

$$= (dY) - \hat{\beta}'(dX) +$$

$$[(dS_{xy}) - \hat{\beta}'(dS_{xx})]S_{xx}^{-1}(X - X).$$

Après quelques manipulations algébriques, nous obtenons:

$$Y^{GREG} - Y = \sum w_k e_k [1 + x_k' S_{xx}^{-1} (X - X)/c_k] + \dots,$$

où  $e_k = y_k - x_k' \hat{\beta}$ . Nous définissons donc:

$$z_k = e_k [1 + x_k' S_{xx}^{-1} (X - X)/c_k].$$

La variance de ce total estimé pour cette variable  $z$  est

identique à la variance proposée par Särndal, Swensson et Wretman (1989). Ces auteurs soutiennent que, en se

basant sur la validité du modèle de régression, cette variance est préférable aux autres estimateurs par développement de Taylor pour le calcul de la variance. Nous constatons que le calcul de cette variable  $z$  se fait naturellement dans notre méthode.

## 2.2 Statistique du test de la somme des rangs de Wilcoxon

Nous démontrons maintenant comment notre méthode

fonctionne dans un cas non standard plus difficile. Nous supposons que nos unités échantillonnées appartiennent à l'une des deux sous-populations, que nous nommons respectivement population 1 et population 2. Nous définissons:

$$I\{x \leq y\} = \begin{cases} 1 & \text{si } x \leq y, \\ 0 & \text{autrement,} \end{cases} \quad \text{et} \quad \delta_k = \begin{cases} 1 & \text{si } k \in \text{Pop. 1} \\ 0 & \text{autrement.} \end{cases}$$

Nous posons:

$$\hat{N}_1(t) = \sum_{k \in s} w_k \delta_k I\{x_k \leq t\},$$

ce qui correspond au nombre estimé d'unités de la population 1 qui ont des valeurs inférieures ou égales à  $t$ . Nous définissons  $\hat{N}_2(t)$  de façon analogue. Notons que  $\hat{N}_j = \hat{N}_j(\infty)$ , le nombre estimé d'unités dans la population  $j$ . Une version pondérée de la statistique du test de Wilcoxon est:

$$T^w = \int_{-\infty}^{\infty} [\hat{N}_1(t) + \hat{N}_2(t)] d\hat{N}_1(t). \quad (13)$$

Cette expression correspond à la somme pondérée des rangs de la population 1 parmi les rangs pondérés de l'échantillon combiné. Afin de calculer la valeur probable asymptotique de  $T^w$  dans (13), posons  $N_i(t) = E[\hat{N}_i(t)]$  pour  $i = 1, 2$ , et substituons  $N_i(t)$  pour  $\hat{N}_i(t)$  dans (13). Nous définissons alors  $F_i(t) = N_i(t)/N_i$ , où  $N_i = E(N_i)$  et nous appliquons l'hypothèse nulle lorsque, disons,  $F_1(t) = F_2(t) = F(t)$ . Nous obtenons ainsi la valeur probable asymptotique suivante:

$$\int_1^1 (N_1 + N_2) F(t) N_1 dF(t) = N_1 (N_1 + N_2) / 2.$$

On notera que dans le cas des échantillons indépendants de taille  $N_1$  et  $N_2$  provenant des populations 1 et 2, respectivement, et en supposant que la fonction de distribution dans chaque population est continue et que les échantillons sont prélevés par échantillonnage aléatoire simple, la valeur probable exacte de  $T^w$  dans (13) est  $N_1(N_1 + N_2 + 1)/2$ .

Considérons maintenant la statistique

$$T^w = \int_{-\infty}^{\infty} [\hat{N}_1(t) + \hat{N}_2(t)] d\hat{N}_1(t) - \frac{N_1(N_1 + N_2)}{2}.$$

Nous utilisons  $\Delta$  plutôt que  $d$  pour dénoter la différentielle totale, car  $d$  est utilisé sous l'intégrale. Nous avons donc

$$(\Delta T^w) = \int_{-\infty}^{\infty} [\Delta \hat{N}_1(t) + \Delta \hat{N}_2(t)] d\hat{N}_1(t)$$

$$+ \int_{-\infty}^{\infty} [\hat{N}_1(t) + \hat{N}_2(t)] d\Delta \hat{N}_1(t)$$

$$- \frac{(\Delta \hat{N}_1)(\hat{N}_1 + \hat{N}_2) + \hat{N}_1(\Delta \hat{N}_1 + \Delta \hat{N}_2)}{2}.$$

En poursuivant selon notre méthode habituelle, nous obtenons:

$$T^w - T^w = \int_{-\infty}^{\infty} \left( \sum w_k I\{x_k \leq t\} \right) d\hat{N}_1(t)$$

$$+ \sum w_k \delta_k [\hat{N}_1(x_k) + \hat{N}_2(x_k)]$$

$$- \frac{\sum w_k \delta_k (\hat{N}_1 + \hat{N}_2) + \hat{N}_1 \sum w_k}{2} + \dots,$$

de sorte que

$$z_k = \sum_j w_j \delta_j I\{x_k \leq x_j\} + \delta_k [\hat{N}_1(x_k) + \hat{N}_2(x_k)]$$

$$- \frac{\delta_k (\hat{N}_1 + \hat{N}_2) + \hat{N}_1}{2}. \quad (14)$$

Nous ne savons pas si ce résultat a déjà été publié. On peut démontrer que lorsque l'hypothèse nulle est vraie et que nous choisissons indépendamment des données dans deux populations par échantillonnage aléatoire simple, et lorsque les populations présentent des fonctions de distribution continues, la variance obtenue à partir des variables  $z$  dans (14) est asymptotiquement équivalente à celle obtenue avec la formule classique habituelle.

$$T = g_1(Y_1, \dots, Y_m, \lambda), \quad \text{ou} \quad \lambda = g_2(Y_1, \dots, Y_m),$$

$$(dT) = \Sigma \left[ \frac{\partial g_1(Y, \lambda)}{\partial Y_i} \right] (dY_i) + \Sigma \left[ \frac{\partial g_1(Y, \lambda)}{\partial \lambda} \right] (d\lambda_j),$$

où

$$(d\lambda_j) = \Sigma^i \left[ \frac{\partial g_{2j}(Y)}{\partial Y_i} \right] (dY_i),$$

$$T - T \equiv \Sigma \frac{\partial g_1(Y, \lambda)}{\partial Y_i} \left( \Sigma^k w_k Y_{ik} - Y_i \right)$$

$$+ \Sigma \frac{\partial g_1(Y, \lambda)}{\partial \lambda_j} \frac{\partial Y_i}{\partial g_{2j}(Y)} \left( \Sigma^k w_k Y_{ik} - Y_i \right)$$

$$= \Sigma w_k z_k + \dots,$$

où

$$z_k = \left[ \frac{\partial g_1(Y, \lambda)}{\partial Y} \right]' Y_k + \left[ \frac{\partial g_1(Y, \lambda)}{\partial \lambda} \right]' \left[ \frac{\partial \lambda}{\partial g_2(Y)} \right] Y_k. \quad (7)$$

Lorsque les paramètres additionnels sont définis seulement de manière implicite par les équations d'estimation, nous obtenons la généralisation suivante:

$$T = g(Y_1, \dots, Y_m, \lambda),$$

où

$$U(Y_1, \dots, Y_m, \lambda) = 0. \quad (8)$$

$$(dT) = \Sigma \left[ \frac{\partial g(Y, \lambda)}{\partial Y_i} \right] (dY_i) + \left[ \frac{\partial \lambda}{\partial g(Y, \lambda)} \right]' (d\lambda),$$

en prenant la différentielle totale de (8) et en isolant  $(d\lambda)$ , nous obtenons:

$$(d\lambda) = - \left[ \frac{\partial U(Y, \lambda)}{\partial \lambda} \right]^{-1} \left[ \frac{\partial U(Y, \lambda)}{\partial Y_i} \right] \Sigma (dY_i). \quad (9)$$

$$T - T \equiv \Sigma^i \left( \frac{\partial g}{\partial Y_i} \right) \left( \Sigma^k w_k Y_{ik} - Y_i \right)$$

$$- \left( \frac{\partial g}{\partial \lambda} \right) \left[ \frac{\partial \lambda}{\partial U} \right] \Sigma^i \left( \frac{\partial Y_i}{\partial U} \right) \left( \Sigma^k w_k Y_{ik} - Y_i \right)$$

$$= \Sigma w_k z_k + \dots,$$

## 2. AUTRES EXEMPLES

Les expressions (6), (7) et (10) ci-dessus sont présentées uniquement afin de donner les formules précises pour les divers cas. Toutefois, dans la pratique, nous recommandons d'utiliser les étapes de base et de partir des principes premiers. Pour démontrer ceci, voici deux exemples. L'un est l'estimateur de régression généralisé (GREG) bien connu; l'autre aboutit à quelques nouveaux résultats pour le test de la somme des rangs de Wilcoxon pour les données des enquêtes complexes.

### 2.1 Estimateur de régression généralisé

On peut écrire comme suit l'estimateur de régression généralisé (GREG) qui est donné, par exemple, dans Särndal, Swensson et Wretman (1989):

$$Y^{GREG} = Y + \beta'(X - X), \quad (11)$$

où le paramètre additionnel  $\beta$  est défini comme étant la solution à:

$$\Sigma^k w_k x_k (Y_k - x_k' \beta) / c_k = 0,$$

et où  $c_k$  est le facteur qui rend la variance hétéroscédastique dans le modèle de régression. Ceci est équivalent à:

$$S_{xx} \beta - S_{xy} = 0, \quad (12)$$

les définitions de  $S_{xx}$  et  $S_{xy}$  étant évidentes. Si l'on prend les différentielles totales dans (12), nous obtenons:

$$(dS_{xx}) \beta + S_{xx} (d\beta) - (dS_{xy}) = 0,$$

de telle sorte que

$$(d\beta) = S_{xx}^{-1} [(dS_{xy}) - (dS_{xx}) \beta].$$

Nous avons donc:

$$\beta - \beta \equiv \Sigma w_k S_{xx}^{-1} [x_k (Y_k - x_k' \beta)] / c_k + \dots$$

Maintenant, en prenant les différentielles totales de (11), nous obtenons:



Nous constatons que cette expression contient un certain nombre d'estimateurs pondérés, ceux qui ont une dépendance explicite par rapport aux valeurs  $w_k$ , ( $\sum w_k y_k$  et  $\sum w_k x_k$ ) et ceux pour lesquels les valeurs  $w_k$  sont implicites dans l'expression ( $\bar{X}$  et  $\bar{R}$ ).

La dernière étape consiste à isoler  $z_k$ , ce que l'on fait en réécrivant l'expression (3) sous la forme:

$$\bar{Y} - \bar{Y} \equiv \sum w_k z_k + \text{autres termes ne dépendant pas explicitement de } w_k.$$

Nous obtenons ainsi:

$$z_k = \frac{\bar{X}}{X} (y_k - R x_k). \quad (4)$$

Nous justifierons à la section 4 la raison pour laquelle

nous ignorons les termes qui ne dépendent pas explicitement de  $w_k$ . On remarquera que  $\sum w_k z_k$  a la forme d'une estimation de la valeur totale de la variable  $z$  pour l'ensemble.

Pour obtenir la variance de  $\bar{Y}_R$ , nous insérons la nouvelle variable  $z_k$  dans l'enregistrement du  $k$ -ième échantillon, et nous utilisons une méthode standard pour calculer la variance d'une valeur totale, appliquée à cette variable. Nous supposons qu'il existe un estimateur de variance possédant de bonnes propriétés pour le plan d'échantillonnage à l'étude.

Nous pouvons résumer comme suit la méthode:

1. Comme estimateur de  $T$ , nous utilisons  $\bar{T}$  et nous prenons sa différentielle totale. Nous supposons que  $\bar{T}$  est asymptotiquement convergent avec le plan.

2. Nous remplaçons la différentielle totale de  $\bar{T}$ ,  $d\bar{T}$ , par  $\bar{T} - T$ . Nous remplaçons toutes les autres différentielles totales des quantités estimées par l'écart par rapport à leurs valeurs probables respectives, comme nous avons remplacé ( $dY$ ) par l'expression ( $\sum w_k y_k - Y$ ), etc.

3. La dernière étape consiste à isoler  $z_k$ , lorsque nous récrivons le résultat de l'étape (2) sous la forme suivante:

$$\bar{T} - T \equiv \sum w_k z_k + \text{autres termes ne dépendant pas explicitement de } w_k.$$

4. Enfin, pour obtenir la variance estimée de  $\bar{T}$ , nous insérons la nouvelle variable  $z_k$  dans chaque enregistrement échantillonné, et nous utilisons la méthode standard (dont on sait qu'elle a de bonnes propriétés) afin de calculer la variance d'un total, appliqué à cette variable.

## 1.1 Cas général le plus simple

Pour les échantillons à une phase, examinons un cas général simple où l'estimateur peut s'exprimer sous la forme d'une fonction différentiable des totaux estimés pour certaines variables de l'enquête, dont certaines peuvent être calculées au niveau de l'unité d'échantillon-nage finale. Dans ce cas-ci, notre méthode donne les résultats suivants:

## 1.2 Cas avec paramètres additionnels

On remarquera que, dans l'expression (6) pour  $z_k$ , tous les termes sont directement obtenus de l'échantillon, de sorte qu'il n'y a pas lieu de substituer des estimateurs pour des quantités inconnues.

Dans de nombreux exemples, on définit plus facilement l'estimateur en termes qui comprennent l'utilisation de paramètres servant uniquement à simplifier la définition du paramètre qui nous intéresse. Pour l'estimateur par quotient,  $R$  est un exemple d'un tel paramètre additionnel. Dans ce cas précis, on dispose d'une équation explicite pour l'estimateur du paramètre additionnel. En présence de paramètres additionnels, on peut écrire comme suit la méthode générale:

$$z_k = \sum^i \left[ \frac{\partial g(Y)}{\partial y_i} \right] y_{ik} = \left[ \frac{\partial g(Y)}{\partial Y} \right]' y_k. \quad (6)$$

où

$$\bar{T} - T \equiv \sum^i \left[ \frac{\partial g(Y)}{\partial y_i} \right] \left( \sum^k w_k y_{ik} - Y_i \right) = \sum w_k z_k + \dots, \quad (5)$$

$$(d\bar{T}) = \sum \left[ \frac{\partial g(Y)}{\partial y_i} \right] (dy_i)$$

$$T = g(Y_1, \dots, Y_m)$$

# Méthodes de linéarisation pour les échantillons à une et deux phases: Une approche de type «recette»

DAVID A. BINDER<sup>1</sup>

## RÉSUMÉ

Il existe un certain nombre de méthodes asymptotiquement équivalentes qui permettent de calculer l'approximation, par série de Taylor, des variances pour les statistiques complexes. Binder et Patak (1994) ont présenté la justification théorique pour une classe de méthodes. Toutefois, on peut établir bon nombre de ces méthodes à partir d'exemples pratiques en utilisant des techniques directes qui ne sont pas décrites clairement dans Binder et Patak. Dans cette communication, nous présentons une approche de type «recette», utilisable pour de nombreux exemples et dont les bonnes propriétés d'échantillonnage de taille finie ont été démontrées. La méthode retenue devient normalement claire lorsque l'on utilise des concepts comme méthodes basées sur un modèle ou linéarisation de la méthode du jackknife; toutefois, notre approche permet d'obtenir les résultats recherchés de façon plus directe. En outre, nous présentons de nouveaux résultats pour l'application de ces techniques à des échantillons à deux phases.

MOTS CLÉS: Enquêtes complexes; estimation de la variance; estimateur par quotient; estimateur par régression; test de la somme des rangs de Wilcoxon; équations d'estimation.

## 1. LA MÉTHODE

Le calcul de la variance asymptotique pour une grande classe d'estimateurs provenant d'échantillons d'enquête complexes est maintenant bien décrit dans la littérature, du moins pour les approximations du premier ordre. Toutefois, il existe un certain nombre d'autres estimateurs de la variance, tous étant asymptotiquement équivalents. Dans la présente communication, nous présentons une méthode simple pour calculer, de façon générale, l'un de ces estimateurs les plus utilisés. Cette méthode simple de calcul peut s'avérer utile pour les statisticiens qui peuvent hésiter devant les choix disponibles et recherchent une solution rapide à leur problème.

Nous débutons par un exemple simple de la méthode utilisant l'estimateur par quotient pour la valeur totale d'un ensemble. Dans ce cas-ci, l'estimateur est:

$$Y_R = R X, \quad (1)$$

pour

$$R = Y/X, \text{ et } Y = \sum_{k \in S} w_k Y_k,$$

où  $s$  est l'ensemble des indices correspondant aux unités échantillonnées et  $w_k$  est le poids d'échantillonnage, normalisé de telle sorte que  $\sum w_k$  soit un estimateur de la valeur totale de l'ensemble; p. ex.,  $w_k = 1/\pi_k$ , où  $\pi_k$  est la probabilité d'inclusion du premier ordre. La définition de  $X$  est analogue à celle de  $Y$ . En prenant la différentielle totale des deux membres de l'expression (1), nous obtenons:

$$(dY_R) = (dR)X, \quad (2a)$$

$$(dR) = \frac{(dY)}{(dY)} - \frac{X}{Y^2} (dX) \quad (2b)$$

$$= \frac{1}{X} [(dY) - R(dX)].$$

Nous constatons que la différentielle totale pour  $\hat{T} = g(Y_1, \dots, Y_m)$  est donnée, en général, par:

$$(d\hat{T}) = \sum \left[ \frac{\partial g(Y)}{\partial Y_i} \right] (dY_i).$$

Nous aurons pu toutefois éviter l'utilisation de  $R$  dans l'expression (1) en définissant tout simplement

$$Y_R = \frac{Y}{X},$$

ce qui élimine la nécessité de définir explicitement  $(dR)$  dans (2b), mais nous l'avons fait afin de rendre plus clairs les exemples complexes qui seront présentés à la section 1.2. Signalons également que l'expression (2a) ne comprend pas la différentielle totale de  $X$ , qui est la valeur totale de la variable  $x$  pour l'ensemble, car  $X$  est supposé être fixe et connu. L'étape suivante consiste à remplacer toutes les différentielles totales des quantités estimées par les écarts par rapport à leurs valeurs probables respectives. Dans le membre de droite, nous remplaçons  $(dY)$  par l'expression  $(\sum w_k Y_k - Y)$ , etc. Pour la quantité qui nous intéresse, soit  $Y_R$ , nous remplaçons  $dY_R$  par  $Y_R - Y$ . À partir de l'expression (2), et faisant ces substitutions, nous obtenons:

$$Y_R - Y = \frac{X}{Y} \left[ \sum w_k Y_k - Y \right] - R \left( \sum w_k X_k - X \right). \quad (3)$$

<sup>1</sup> David A. Binder, Directeur, Division des méthodes d'enquêtes-entreprises, Statistique Canada, R.H. Coats, 11 A, Ottawa (Ontario), Canada, K1A 0T6.





- CLARK, W.A.V., et AVERY, K.L. (1976). The effect of data aggregation in statistical analysis. *Geographical Analysis*, 8, 428-438.
- DUNCAN, D.P., et DAVIS, B. (1953). An alternative to ecological correlation. *American Sociological Review*, 18, 665-666.
- FOTHERINGHAM, A.S., et WONG, D.W.S. (1991). The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning, A*, 23, 1025-1044.
- GEHLKE, C.E., et BIEHL, K. (1934). Certain effects of grouping upon the size of the correlation coefficient in census tract material. *Journal of the American Statistical Association*, 29, Supplement, 169-170.
- GOODMAN, L.A. (1959). Some alternatives to ecological correlation. *American Journal of Sociology*, 64, 610-625.
- HANNAN, M.T., et BUSTEIN, L. (1974). Estimation from grouped observations. *American Sociological Review*, 39, 374-392.
- HOLT, D., SMITH, T.M.F., et WINTER, P.D. (1980). Regression analysis of data from complex surveys. *Journal of the Royal Statistical Society, A*, 143, 474-87.
- HOLT, D., et SCOTT, A.J. (1981). Regression analysis using survey data. *The Statistician*, 30, 169-173.
- LICHTMAN, A.J. (1974). Correlation, regression, and the ecological fallacy: A critique. *Journal of Interdisciplinary History*, 4, 417-433.
- LANGBEIN, L.I., et LICHTMAN, A.J. (1978). *Ecological Inference*. Thousand Oaks, CA: Sage.
- OPENSHAW, S. (1984). Ecological fallacies and the analysis of areal census data. *Environment and Planning, A*, 6, 17-31.
- OPENSHAW, S., et TAYLOR, P.J. (1979). A million or so areal unit problem. Dans *Statistical Applications in the Spatial Sciences*, (Ed. N. Wrigley), 127-144.
- PEARSON, K. (1903). On the influence of natural selection on the variability and correlation of organs. *Philological Transactions of the Royal Society, A*, 200, 1-66.
- PERLE, E.D. (1977). Scale changes and impacts on factorial ecology structures. *Environment and Planning, A*, 9, 549-558.
- RAO, C.R. (1973). *Linear Statistical Inference and its Applications*, (2<sup>ème</sup> éd.). New York: Wiley.
- ROBINSON, W.S. (1950). Ecological correlations and the behaviour of individuals. *American Sociological Review*, 15, 351-357.
- SMITH, K.W. (1977). Another look at the clustering perspective on aggregation problems. *Sociological Methods and Research*, 5, 289-316.
- SMITH, T.M.F., et HOLMES, D. (1989). Multivariate analysis. Dans *Analysis of Complex Surveys*, (Eds. C.J. Skinner, D. Holt et T.M.F. Smith), 165-187.
- STEELE, D. (1985). Statistical Analysis of Populations with Group Structure. Unpublished PhD Thesis, Department of Social Statistics, University of Southampton.
- STEELE, D., et HOLT, D. (1994). Analysing and Adjusting Aggregation Effects: The Ecological Fallacy Revisited. Department of Applied Statistics, University of Wollongong, préirage 1/94.
- STEELE, D., et HOLT, D. (1995). Rules for random aggregation. *Environment and Planning (à paraître)*.
- YULE, U., et KENDALL, M.S. (1950). *An Introduction to the Theory of Statistics*. Glendale, CA: Griffin.



et  $r_{ab}$  (0,186) donne une indication supplémentaire des conséquences de la correction. Après correction au moyen des quatre variables, l'écart n'est plus que de 0,126 (0,090 après correction au moyen de sept variables). Les valeurs médianes correspondantes pour  $|s_{ab} - s_{ab}|$  sont respectivement de 0,173, de 0,039 et de 0,017.

## 5. CONCLUSIONS ET DISCUSSION

Les auteurs proposent un modèle qui décompose en deux éléments le biais observé dans l'analyse au niveau du groupe au moyen des matrices des covariances pour des populations groupées. Le premier élément résulte des variables de groupement et le second, des corrélations résiduelles entre les variables  $y$  à l'intérieur du groupe, compte tenu des variables de groupement  $z$ . Pareille décomposition nous aide à comprendre l'ampleur des effets cumulatifs. Elle indique aussi comment supprimer le biais attribuable aux variables de groupement quand on dispose de renseignements supplémentaires sur la matrice des covariances de niveau unitaire des variables de groupement.

Beaucoup de données de groupe à divers degrés d'agrégation peuvent être extraites du recensement et de nombreuses autres sources dans maints pays. L'expansion des systèmes d'information géographique accroîtra l'accessibilité de telles données. Il est important d'analyser et de décomposer les effets de groupe. La théorie et la méthode exposées ici procurent le cadre nécessaire à un tel exercice. En identifiant les variables qui expliquent la majeure partie des effets de groupe, donc qui devraient permettre la correction des analyses écologiques, on ouvrira la voie à l'utilisation des données agrégées.

## REMERCIEMENTS

Le présent projet de recherche n'aurait pu être mené à bien sans la subvention H507 26 5013 de l'Economic and Social Research Council du Royaume-Uni. Les auteurs remercient sincèrement les examinateurs et le rédacteur adjoint pour leurs commentaires utiles.

## BIBLIOGRAPHIE

- ARBIA, G. (1989). *Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems*. Dordrecht: Kluwer.
- BLALOCK, H.M. (1964). *Causal Inference in Nonexperimental Research*. Chapel Hill NC: University of North Carolina Press.
- BLALOCK, H.M. (1979). Measurement and conceptualization problems: The major obstacle to integrating theory and research. *American Sociological Review*, 44, 881-894.
- BLALOCK, H.M. (1985). Cross level analysis. Dans *The Collection and Analysis of Community Data*, (Ed. J.B. Casterlin), ISI, World Fertility Survey.

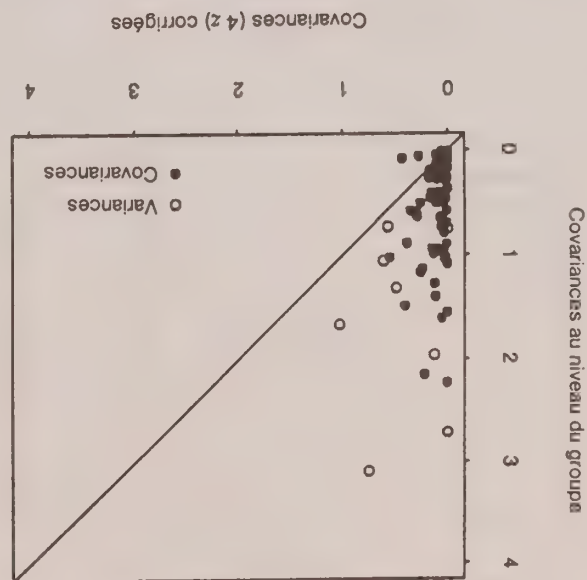


Figure 3a.

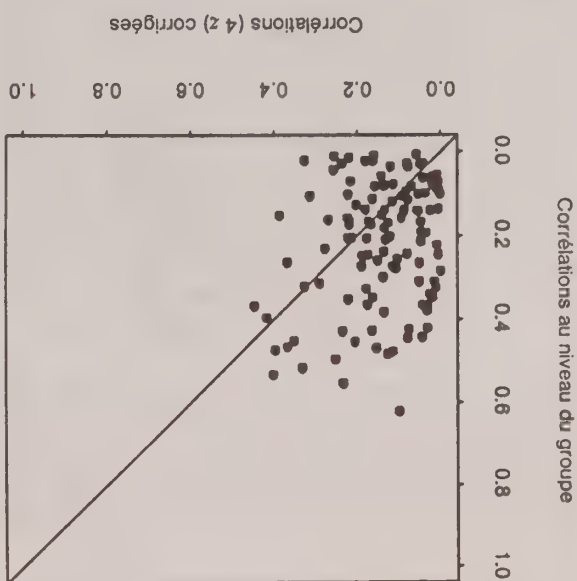


Figure 3b.

covariances de départ de la matrice  $16 \times 16$  doivent donc être corrigées au moyen des 21 variances et covariances des variables  $z$ . Pour avoir une idée des résultats qu'on pourrait obtenir avec un minimum d'information, nous avons restreint les variables d'ajustement aux quatre variables liées à l'âge et au mode d'occupation. Le tableau 4 indique que ces variables expliquent respectivement 57% et 71% des deux mesures de l'agrégation. Les figures 3a et 3b donnent le tracé correspondant aux figures 2a et 2b pour ce cas particulier. La figure 1c compare les corrélations corrigées au moyen des quatre variables aux corrélations de niveau individuel. Bien sûr, la correction n'est pas aussi efficace, mais il est encourageant de voir ce qu'on obtient avec si peu de variables. L'écart médian absolu entre  $r_{ab}$





Tableau 2

Cinq premières VGC des variables du tableau 1

	VG1	VG2	VG3	VG4	VG5
18 à 29 ans	0.4	0.3	0.9	1.1	0.1
30 à 44 ans	0.1	0.5	0.36	1.0	0.2
45 à 59 ans*	-0.1	1.2	-0.2	1.0	0.1
60 ans et plus*	0.3	2.2	-0.5	2.6	0.9
Femme	0.1	0.0	0.0	0.3	0.1
Personne de couleur*	0.5	-0.4	1.4	-1.1	5.2
Marié	-0.2	-0.5	-0.4	-0.8	-0.1
Maladie débilitante à long terme	0.3	0.1	-0.2	0.2	0.3
Travailleur à temps plein	0.7	-0.3	0.2	1.2	0.4
Chômeur	0.7	0.0	-0.1	0.0	-0.4
Autre situation d'emploi	0.1	0.1	0.0	-0.2	-0.1
Soutien de famille né au R.-U.	0.5	-0.1	-1.0	0.4	0.2
Soutien de famille né dans le Nouveau Commonwealth	0.0	-0.1	-0.3	0.1	0.6
Soutien de famille émigré ≤ 0.5 personne par pièce	0.2	0.1	1.4	0.6	-1.3
Personnes dans un ménage sans automobile	2.2	0.6	0.8	-1.9	-0.7

\* Variable retenue pour la correction.  
Source: Reigate et Banstead, données du recensement de 1991 pour le DL de Tandridge.

Connaissant  $S_{yy}$  et  $S_{yys_1}$ , on peut analyser les variables de groupement canoniques en vue de comprendre la structure particulière du groupe. Le tableau 2 donne la charge des cinq premières variables de groupement canoniques. Ces cinq variables expliquent 89% de l'effet cumulatif des 16 variables utilisées.

La première VGC influe fortement sur la densité de population dans le logement et sur l'accès à une automobile. On pourrait donc la considérer comme un paramètre socio-économique. La deuxième attribue une forte charge aux variables qui identifient les personnes des deux groupes les plus âgés. La contribution des chefs de famille de couleur à la variable suivante est également remarquable. Comme on pouvait s'y attendre, certaines variables comme la proportion de femmes ne présentent pratiquement aucune corrélation à l'intérieur du groupe, donc n'ont aucun effet cumulatif et n'exercent presque aucune influence sur les VGC. Pareilles variables ne changent pas d'une région à l'autre. Elles n'ont donc habituellement aucune valeur explicative.

Dans la pratique, pareille analyse des VGC serait irréalisable, l'existence même de  $S_{yy}$  rendant une analyse agrégée inutile. L'analyse des VGC a néanmoins le mérite d'identifier les variables les plus importantes en raison de la charge qu'elles imposent aux premières VGC. On sait pertinemment qu'au Royaume-Uni, les variables associées au mode d'habitation (qui ne font pas partie des 16 variables intéressantes) partagent des liens très étroits avec maintes variables liées à la situation socio-économique, au comportement et à la santé. Il y a tout lieu de croire qu'en les utilisant comme variables auxiliaires 2 pour la correction, on tiendrait compte d'une bonne partie de la première dimension socioéconomique. Elles pourraient donc remplacer les variables illustrant la densité de la population dans le logement et l'accès à une automobile,

$$\|S_{yys_1} - \bar{S}_{yy}\| - \|S_{yys_1} - \bar{S}_{yy}(z)\|$$

qui est la réduction de l'effet agrégé multivarié. La seconde est

$$1 - \frac{\text{tr}(S_{yys_1}^{-1} \bar{S}_{yy})}{\text{tr}(S_{yys_1}^{-1} \bar{S}_{yy}(z))} - 1$$

l'ajustement. La première est donnée par

Source: Reigate et Banstead, données du recensement de 1991 pour le DL de Tandridge.

Variable	Mode	d'occupation	Type	Corrélation
Effet d'agrégation	Locataire	Propriétaire occupant	Unif./semi/terrasse	
Effet d'agrégation	0.261	133.43	90.03	0.175
Corrélation	0.177	90.83	59.52	0.113

Tableau 3

Effet d'agrégation et corrélation à l'intérieur des variables liées au ménage dans le DL de Reigate

soit les sept variables d'ajustement potentielles qui suivent, sous les trois variables d'intérêt identifiées au tableau 1 par un astérisque (45-59 ans, 60 ans +, personnes de couleur) et les quatre variables liées au logement du tableau 3, avec leurs effets cumulatifs et leurs corrélations à l'intérieur de la grappe.

qu'on estime influencer fortement sur la première VGC. L'autre raison pour envisager l'utilisation de ces variables est que si l'analyse actuelle doit illustrer ce qui pourrait se produire dans d'autres situations, les variables de base sur le mode d'occupation et sur le logement seront sans doute plus faciles à obtenir que celles sur la densité de la population dans le logement et l'accès à une automobile pour la correction. Compte tenu des résultats de l'analyse des VGC et puisqu'on souhaite des variables d'ajustement faciles à retrouver, peu importe la situation, nous proposons les sept variables d'ajustement potentielles qui suivent, sous les trois variables d'intérêt identifiées au tableau 1 par un astérisque (45-59 ans, 60 ans +, personnes de couleur) et les quatre variables liées au logement du tableau 3, avec leurs effets cumulatifs et leurs corrélations à l'intérieur de la grappe.

Le tableau 4 illustre l'effet de diverses combinaisons de variables sur la correction des résultats de l'analyse agrégée. Les deux variables associées à l'âge présentent manifestement de l'importance (elles expliquent 38% de l'effet d'agrégation multivariée et 53% de l'écart généralisé), mais il en va autant des variables sur le mode

**Tableau 1**  
Effet d'agrégation et corrélation à l'intérieur de la classe pour les variables du recensement dans le DL de Reigate

Effet d'agrégation	Corrélation	Variable retenue pour la correction.
		* Variable retenue pour la correction.
		Source: Reigate et Banstead; données du recensement de 1991 pour le DL de Tandridge.
		DL de Tandridge.
		défini pour les éléments diagonaux appropriés de $\Delta_{yy}$ et de $\Sigma_{yy}$ , correspond à la corrélation de la $a$ -ième variable à l'intérieur du groupe. On peut obtenir l'estimation $\delta_{aa}$ de la corrélation à l'intérieur du groupe à partir de (2.18), puisque $\bar{Q}_a = 1 + (\pi^* - 1) \delta_{aa}$ . Les résultats relatifs à ces variables apparaissent au tableau 1. En général, les corrélations à l'intérieur du groupe sont faibles, mais étant donné le nombre d'observations pour chaque DR, les effets cumulatifs peuvent être importants (voir la remarque qui suit l'équation (2.18)).

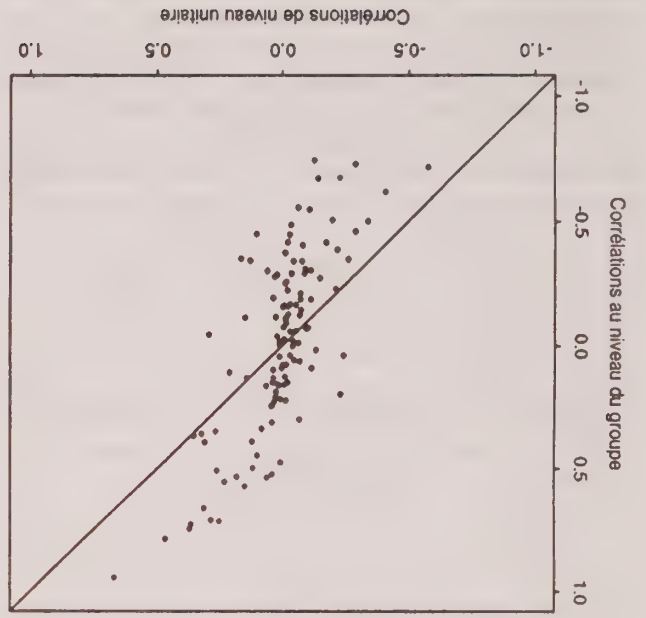


Figure 1a.

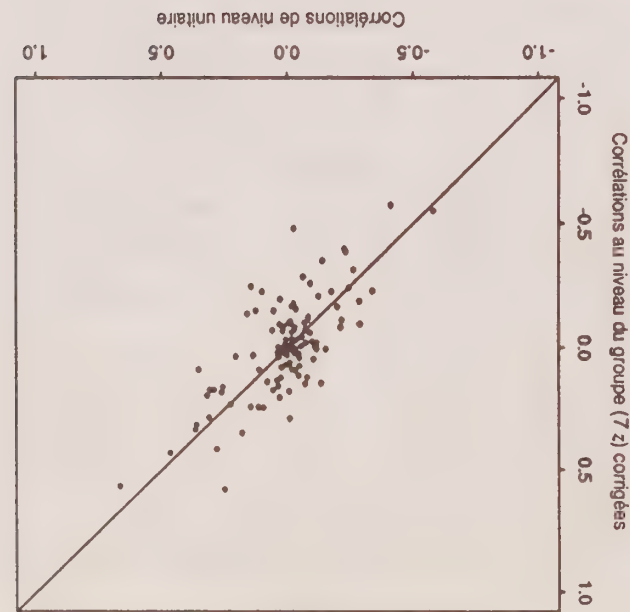


Figure 1b.

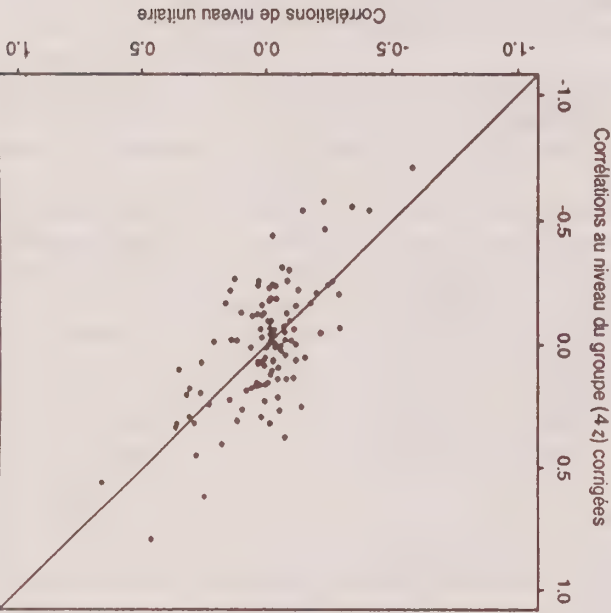


Figure 1c.

La figure 1a montre la courbe de la corrélation au niveau du groupe  $\bar{r}_{ab}$  par rapport à la corrélation de niveau individuel  $r_{ab}$  pour chaque paire de variables. La forme en S caractéristique de la courbe fait ressortir l'importance des effets dus à l'agrégation. Les petites corrélations de niveau unitaire sont habituellement amplifiées si bien que dans la plupart des cas,  $|r_{ab}|$  est beaucoup plus grand que  $|\bar{r}_{ab}|$ .



$$\bar{\Sigma}_{yy}(a^q) = S_{yys_1} + \sum_p^k (\theta_k - 1) \phi_k \phi_k$$

et  $\text{tr}(S_{yys_1}^{-1} \bar{\Sigma}_{yy}(a^q)) = \sum_{k=q+1}^p \theta_k$ . En fait, quand on prend les  $q$  premières VGC, on obtient la matrice de rang  $q$  qui minimise  $\|S_{yys_1} - \bar{\Sigma}_{yy}(a^q)\|$ . Par conséquent, en étudiant les quantités

$$\sum_p^k \theta_k \quad \text{et} \quad 1 + \sum_p^k (\theta_k - 1) \hat{\rho}_{2ak}^{2k}$$

$$\text{pour } q = 0, \dots, p - 1$$

il est possible d'établir comment les  $q$  premières VGC de l'échantillon expliqueront la partie de l'effet d'agrégation global et l'effet d'agrégation individuel de chaque variable. Une telle analyse indiquera combien de dimensions sont nécessaires pour expliquer les effets cumulatifs et en supprimer une partie précise. Par ailleurs, en examinant la charge des variables à l'origine des VGC, on devrait être en mesure d'identifier les variables qui «expliqueront» le mieux les effets cumulatifs des autres variables. Les chercheurs devraient s'efforcer d'obtenir les données de niveau unitaire sur ces variables afin de les appliquer à l'estimateur corrigé.

Les résultats qui précèdent ont certaines implications importantes quant à l'usage des données au niveau du groupe complètes par des données de niveau unitaire limitées. En effet, ils autorisent la combinaison des données d'enquête et des données de groupe d'une ou de plusieurs sources, et laissent entrevoir une stratégie d'analyse pour les effets de groupe et les données au niveau du groupe.

#### 4. QUELQUES RÉSULTATS EMPIRIQUES

Nous illustrerons les principes énoncés précédemment par une analyse des données du recensement de la population du Royaume-Uni effectuée en 1991 pour le district local (DL) de Reigate, Banstead et Tandridge. Ce DL compte 188,700 habitants répartis dans 371 DR, ce qui donne une population moyenne de  $n = 508.6$  par DR. On possède les données de groupe pour les habitants de chaque DR grâce au fichier de données régionales (FDR). Les données de niveau unitaire correspondantes du DL viennent d'un échantillon de 2 pour cent d'enregistrements sur des sujets anonymes. Il est donc impossible d'associer un enregistré à un DR précis, de telle sorte que  $S_{yy}$  repose sur les données complètes du fichier se rapportant au DR et que l'estimation de  $S_{yys_1}$  repose sur l'échantillon de 2 pour cent d'enregistrements anonymes. L'analyse qui suit tient compte de 16 variables de recensement, pour chaque personne.

On s'est servi des données de groupe et des données de niveau unitaire de chaque variable pour calculer l'effet cumulatif  $\bar{Q}_a = s_{aa}/s_{aa}$ . Le paramètre  $\delta_{aa} = \Delta_{aa}/\Sigma_{aa}$ ,

Une fois les VGC calculées, il est possible d'exprimer l'écart entre la matrice des covariances de l'échantillon au niveau du groupe et au niveau unitaire de la façon suivante:

$$S_{yy} - S_{yys_1} = \sum_k^k (\theta_k - 1) \phi_k \phi_k$$

où  $\phi_k$  correspond au vecteur des covariances de l'échantillon entre la  $k$ -ième VGC et les variables originales. On peut donc diviser l'écart entre la matrice des covariances du niveau du groupe et celle du niveau unitaire en  $k$  composantes orthogonales, soit une pour chaque VGC. En ce qui concerne la covariance entre  $y_{ia}$  et  $y_{ib}$ , l'écart entre la covariance de l'échantillon au niveau du groupe  $s_{ab}$  et celle du niveau unitaire  $s_{ab}$  ( $s_{ab}$  étant respectivement des éléments de  $S_{yy}$  et de  $S_{yys_1}$ ) est donnée par

$$s_{ab} = s_{ab} + (s_{aa}s_{bb})^{1/2} \sum_k^k (\theta_k - 1) \hat{\rho}_{ak} \hat{\rho}_{bk}$$

où  $\hat{\rho}_{ak} = \phi_{ak}/s_{aa}^{1/2}$  représente la corrélation pour la  $a$ -ième variable et la  $k$ -ième VGC de l'échantillon. En utilisant les  $q$  premières VGC de l'échantillon pour créer une matrice des variances corrigée au niveau du groupe, c'est-à-dire  $\hat{a}_{qi} = \hat{D}_q^q y_i$  où  $\hat{D}_q = [d_1, \dots, d_q]$  servent de variables auxiliaires,

$$\bar{\Sigma}_{yy}(a^q) = S_{yy} + B_{y_{nq}}' (S_{nq_{nq}so} - S_{nq_{nq}}) B_{y_{nq}}$$

on élimine les  $q$  premiers termes de la décomposition, donc

### 3. IDENTIFICATION DES VARIABLES DE GROUPEMENT

Dans la partie qui précède, nous avons présenté un ensemble de variables auxiliaires 2 caractérisant les variations régionales. Il a servi à corriger l'analyse agrégée et à réduire le biais attribuable à l'aggrégation. Avec des variables auxiliaires idéales,  $\Delta_{yy,z}$  serait égal à zéro et la méthode de correction supprimerait totalement le biais dû à l'aggrégation. En pratique cependant, on ignore quelles variables auxiliaires permettraient d'obtenir  $\Delta_{yy,z} = 0$ . C'est pourquoi nous devons nous en tenir à des variables pour lesquelles on connaît la moyenne régionale, compte tenu des données analysées, et pour lesquelles il est possible d'estimer la matrice des covariances au niveau unitaire  $\Sigma_{zz}$ . On pourra recourir notamment aux données démographiques et aux variables sur l'habitation de base que fournit couramment le recensement. Toutefois, il se peut que ces variables ne caractérisent pas exactement le processus de groupement, donc n'expliquent pas les différences entre régions aussi bien qu'on le souhaiterait.

Dans la partie qui précède, nous avons présenté un ensemble de variables auxiliaires 2 caractérisant les variations régionales. Il a servi à corriger l'analyse agrégée et à réduire le biais attribuable à l'aggrégation. Avec des variables auxiliaires idéales,  $\Delta_{yy,z}$  serait égal à zéro et la méthode de correction supprimerait totalement le biais dû à l'aggrégation. En pratique cependant, on ignore quelles variables auxiliaires permettraient d'obtenir  $\Delta_{yy,z} = 0$ . C'est pourquoi nous devons nous en tenir à des variables pour lesquelles on connaît la moyenne régionale, compte tenu des données analysées, et pour lesquelles il est possible d'estimer la matrice des covariances au niveau unitaire  $\Sigma_{zz}$ . On pourra recourir notamment aux données démographiques et aux variables sur l'habitation de base que fournit couramment le recensement. Toutefois, il se peut que ces variables ne caractérisent pas exactement le processus de groupement, donc n'expliquent pas les différences entre régions aussi bien qu'on le souhaiterait.

#### 3.1 Stratégie d'analyse

Dans la pratique, les variables de groupement ne sont pas connues. Il faut donc élaborer une stratégie en vue d'identifier les variables d'ajustement pour lesquelles on possède une estimation de la matrice des covariances au niveau unitaire, et qui expliqueront les effets de groupe-ment. Voici une façon d'y arriver :

1) Identifier un ensemble de variables couvrant le même domaine que les variables auxquelles on s'intéresse, mais pour lesquelles il existe des données de niveau régional et de niveau unitaire pour une période quelconque du passé (un recensement antérieur, par exemple).

2) Ajouter aux variables qui précèdent d'autres variables (variables démographiques et variables sur l'habitation, par exemple) susceptibles de remplacer les variables 2 qui présentent une étroite association avec les variations régionales. On aura aussi besoin d'une estimation des matrices des covariances de niveau régional et de niveau unitaire pour la même période.

3) Analyser les données pour déterminer les variables qui expliquent le mieux les effets de niveau régional sur les variables d'intérêt. Nous reviendrons plus tard à cette analyse, baptisée analyse des VGC.

4) De (3), trouver la série de variables de correction qu'on pourrait extraire de l'ensemble de données courant et pour lesquelles on pourrait obtenir la matrice des covariances courante de niveau unitaire d'une source quelconque.

5) Il est possible qu'on découvre une estimation de la variance ou de la covariance de niveau unitaire pour certaines variables qui nous intéressent dans les tables existantes, par exemple. Ensuite, il faut encore calculer les effets de l'aggrégation  $\bar{Q}_{aa} = \bar{s}_{aa}/s_{aa}$  ou  $\bar{Q}_{ab} = \bar{s}_{ab}/s_{ab}$ .

#### 3.2 Variables de groupement idéales

Nous examinerons d'abord l'ensemble idéal de variables de groupement utilisables pour la correction de façon à déterminer quelle analyse (VGC) convient le mieux aux données agrégées, conformément à la démarche décrite plus haut.

Supposons que pour l'ensemble complet des variables covariances de niveau régional  $\Sigma_{yy}$  et la matrice des variances-covariances de niveau unitaire  $\Sigma_{yy^{s1}}$  pour l'échantillon  $s_1$ . Bien sûr, si la chose était faisable dans la réalité, le problème d'aggrégation ne se poserait pas puisqu'on pourrait écartier  $\Sigma_{yy}$  et simplement prendre  $\Sigma_{yy^{s1}}$  comme estimation de  $\Sigma_{yy}$ . Néanmoins, il y a trois bonnes raisons pour étudier pareille situation. Tout d'abord, il vaut la peine d'éclaircir la structure du groupement qui établit la relation entre  $\Sigma_{yy}$  et  $\Sigma_{yy^{s1}}$ . En second lieu, il se pourrait que  $\Sigma_{yy}$  et  $\Sigma_{yy^{s1}}$  soient connus pour un moment quelconque dans le temps, le jour du recensement par exemple, mais que l'analyse approfondie d'une version plus récente de  $\Sigma_{yy}$  doive reposer sur des données intercenitaires pour lesquelles on ignore  $\Sigma_{yy^{s1}}$ . Si la structure du groupement persiste dans le temps, comme il est raisonnable de le penser, l'analyse de  $\Sigma_{yy}$  et de  $\Sigma_{yy^{s1}}$  au jour du recensement pourrait faciliter l'analyse intercenitaire en permettant l'identification des variables clés qui expliquent la majeure partie des effets de l'aggrégation. Ces possibilités sont à la base même de la stratégie décrite à la partie 3.1. Troisièmement, si les variables de  $y$  couvrent de nombreuses variables socio-économiques et démographiques, comme cela se produit avec le recensement, les variables clés à l'origine des effets de groupement des variables à l'étude pourraient aussi expliquer une bonne partie des effets de groupement des autres variables socio-économiques et démographiques. Soulignons que les deux échantillons  $s$  et  $s_1$  peuvent être identiques, sans que cela soit une condition. Ainsi,  $s$  pourrait venir d'une source administrative, en fait une sorte de recensement fournissant des données agrégées sur les régions, tandis que  $s_1$  pourrait comprendre les données individuelles d'un sondage sans marqueur géographique. Pour faciliter l'identification des variables importantes associées au groupement, Steel (1985) suggère de calculer les valeurs propres  $\theta_1, \dots, \theta_p$  de  $\Sigma_{yy^{s1}} \Sigma_{yy}^{-1}$  et la matrice  $\bar{D}_y = [\bar{d}_1, \dots, \bar{d}_p]$  de telle sorte que

$$\bar{D}_y \bar{S}_{yy} \bar{D}_y = \text{diag}(\theta_k) \quad \text{et} \quad \bar{D}_y' \bar{S}_{yy^{s1}} \bar{D}_y = I.$$

Les variables définies par la transformation

$$u_i = \bar{D}_y' y_i$$

ont un ratio maximal pour la variance des groupes par rapport à l'échantillon, une corrélation nulle pour



Par ailleurs, selon Steel (1985), (2.36) et (2.38) impliquent que

$$E[\mathcal{L}_{yy}(z) | s, z, s_0, c] = \mathcal{L}_{yy} + \beta_{yz}'(S^{zzs_0} - \mathcal{L}_{zz})\beta_{yz}$$

$$+ (\pi^* - 1)\Delta_{yy,z} + 0(m^{-1}) \quad (2.42)$$

pourvu que  $\text{tr}(S^{z_0}_{-1} S^{zzs_0})$  et  $n \text{tr}((S^{z_0}_{-1} S^{zzs_0} - I)S^{z_0}_{-1} S^{zz})$  soient bornés,  $S^{zz}_2$  étant défini de la même façon que  $S^{zz}$ , sauf que  $n_g$  est remplacé par  $n_g^2/n$ .

Quand on compare (2.42) à (2.35), on se rend compte que la composante du biais attribuable aux variables de regroupement a été ramenée au biais lié à l'utilisation de  $S^{z_0}_{-1}$  si elle est disponible. L'estimateur corrige les effets cumulatifs résultant de  $z$  et ramène l'effet du plan d'échantillonnage associé à  $s$  à celui lié à  $s_0$ .

Supposons que le plan d'échantillonnage utilisé pour produire  $s_0$  et les valeurs des variables auxiliaires viennent d'une superpopulation, de telle sorte que

$$E[\bar{z}_{s_0} | s_0, c] = \mu_z + 0(m_0^{-1}) \quad (2.43)$$

$$E[S^{zzs_0} | s_0, c] = \mathcal{L}_{zz} + 0(m_0^{-1}) \quad (2.44)$$

où  $m_0$  représente le nombre de groupes dans  $s_0$ .

Alors,

$$E[\mu_y(z) | s, s_0, c] = \mu_y + 0(m_0^{-1}) \quad (2.45)$$

$$E[\mathcal{L}_{yy}(z) | s, s_0, c] = \mathcal{L}_{yy} + (\pi^* - 1)\Delta_{yy,z} + 0(m^{-1}) \quad (2.46)$$

ou

$$\bar{m} = \min(m, m_0).$$

Les conditions (2.43) et (2.44) s'appliquent si les valeurs

des composantes de la variance sont similaires à celles du modèle A et si le plan d'échantillonnage de  $s_0$  ne dépend que du regroupement, pas des variables auxiliaires. L'échantillonnage aléatoire simple, l'échantillonnage par grappes à probabilité de sélection identique, voire l'échantillonnage à plusieurs degrés respectent cette condition. On pourrait aussi se servir des données du recensement pour que  $s_0$  constitue la population finie dans son ensemble.

Le biais attribuable aux variables de regroupement peut donc être corrigé, pourvu qu'on dispose d'une matrice des covariances quelconque pour  $z$  au niveau unitaire. La raison pour laquelle on recourt à une telle approche est de créer une situation où les effets prédominants du groupe seraient attribués à la sélectivité ou au regroupement, par le truchement des variables de regroupement. La correction relative aux variables auxiliaires supprime l'effet de la corrélation apparente à l'intérieur du groupe associée à ces variables. Même corrigé cependant, l'estimateur inclut un élément de biais en raison de  $\Delta_{yy,z}$  et si  $z$  n'atténue pas assez les corrélations à l'intérieur du groupe, le biais demeure appréciable. Cette approche dépend donc du choix des variables auxiliaires qui réduiront les corrélations à l'intérieur du groupe.

Si le plan d'échantillonnage de  $s_0$  et le modèle de superpopulation de  $z$  interdisent l'application de (2.43) et de (2.44), on peut remplacer  $\bar{\mu}_{z_0}$  et  $S^{zzs_0}$  par les estimateurs  $\hat{\mu}_{z_0}$  et  $\hat{\mathcal{L}}_{zzs_0}$  pour obtenir des estimateurs corrigés  $\hat{\mu}_y(z)$  et  $\hat{\mathcal{L}}_{yy}(z)$ .

L'espérance mathématique des estimateurs corrigés est donnée par (2.41) et (2.42), où il suffit de remplacer  $\bar{\mu}_{z_0}$  par  $\hat{\mu}_{z_0}$  et  $S^{zzs_0}$  par  $\hat{\mathcal{L}}_{zzs_0}$ . Plusieurs possibilités peuvent être retenues pour  $\hat{\mu}_{z_0}$  et  $\hat{\mathcal{L}}_{zzs_0}$  à partir de l'échantillon  $s_0$ . Smith et Holmes (1989) ont examiné toute une gamme d'estimateurs envisageables, en fonction du modèle et du plan d'échantillonnage. Supposons, par exemple, que le plan d'échantillonnage utilisé pour obtenir  $s_0$  comporte une stratification d'après les valeurs du vecteur des variables de taille  $x$ . Appelons la probabilité d'inclusion de l'unité de population  $i$  dans l'échantillon,  $\Pi_i$ . Le poids associé à cette probabilité serait  $w_i = (\Pi_i)^{-1}$ . L'estimateur de  $\mu_z$  pondéré pour la probabilité est  $\bar{\mu}_{z_0} = \sum_{i \in s_0} w_i z_i$ , et celui de  $\mathcal{L}_{zz}$  est  $\mathcal{L}^{zzs_0} = \sum_{i \in s_0} w_i z_i z_i' - w_0^{-1} \bar{\mu}_{z_0} \bar{\mu}_{z_0}'$  où  $w_0 = \sum_{i \in s_0} w_i$ .

Les estimateurs corrigés de Pearson pour  $\mu_z$  et  $\mathcal{L}_{zz}$  sont  $\bar{z}_{s_0} + B'_{z_0}(\bar{x}_n - \bar{x}_{s_0})$  et  $S^{zzs_0} + B'_{z_0}(S^{zzx_n} - S^{zzx_{s_0}})$  respectivement. Dans ce cas,  $\bar{x}_n$  et  $S^{zzx_n}$  correspondent au vecteur des moyennes et à la matrice des covariances des variables du plan d'échantillonnage de  $x$  pour la population finie, et  $B_{z_0} = S^{zx_{s_0}-1} S^{zzs_0}$ .

On pourrait aussi utiliser les estimateurs de Pearson corrigés et pondérés pour la probabilité, à savoir  $\bar{z}_{s_0}^*$  et  $B'_{z_0}(\bar{x}_n - \bar{x}_{s_0}^*)$  et  $S^{zzs_0} + B'_{z_0}(S^{zzx_n} - S^{zzx_{s_0}^*})$ . Ici  $\bar{x}_{s_0}^*$  et  $S^{zzx_{s_0}^*}$  correspondent respectivement à  $\bar{z}_{s_0}^*$  et  $S^{zzs_0}$  et où  $B_{z_0}^{zzs_0} = S^{zx_{s_0}-1} S^{zzs_0}$ . Puisque, jusqu'à présent, l'approche repose essentiellement sur l'utilisation d'un modèle, il serait préférable de prendre les estimateurs  $\mu_y$  et  $\mathcal{L}_{zz}$ . En pratique néanmoins, les données dont on pourrait se servir pour la correction comprendraient les estimateurs connus de la moyenne et de la covariance, pondérés pour la probabilité et tirés de l'échantillon  $s_0$ , qui est indépendant de  $s$ . Par conséquent,

$$E_{p_0}[\hat{\mu}_{z_0} | z, c] = \bar{z}_n$$

$$E_{p_0}[\hat{\mathcal{L}}_{zzs_0} | z, c] = S^{zzn}$$

où  $\bar{z}_n$  et  $S^{zzn}$  correspondent au vecteur des moyennes et à la matrice des covariances des variables auxiliaires de la population finie, et où  $E_{p_0}$  représente l'espérance mathématique d'un échantillon répété avec le plan d'échantillonnage utilisé pour obtenir  $s_0$ , c'est-à-dire la distribution de randomisation. De (2.41) et (2.42), on peut déduire que

$$E[\hat{\mu}_y(z) | s, z, c] = \mu_y + \beta_{yz}'(\bar{z}_n - \mu_z)$$

$$E[\mathcal{L}_{yy}(z) | s, z, c] = \mathcal{L}_{yy} + \beta_{yz}'(S^{zzn} - \mathcal{L}_{zz})\beta_{yz}$$

$$+ (\pi^* - 1)\Delta_{yy,z} + 0(m^{-1}).$$

Ces espérances mathématiques sont reprises dans le modèle statistique qui génère les valeurs  $y$  et la répartition par randomisation associée à  $s_0$ . En réalité  $\bar{z}_n$  et  $S^{zzn}$  se rapprochent beaucoup de  $\mu_z$  et  $\mathcal{L}_{zz}$  respectivement.

La matrice  $S_{yy}^z$  pondérée au niveau du groupe permet d'estimer  $\Sigma_{yy}^z$ . Le premier terme de biais de (2.36) vient de l'effet des variables de groupement. Il est égal à zéro si  $\beta_{yz} = 0$ , ou presque égal à zéro si  $S_{zz}^z \approx \Sigma_{zz}^z$ . La condition  $\beta_{yz} = 0$  est ferme et signifie que les variables intéressantes n'ont aucun lien avec les variables de groupement. L'effet de l'agrégation sur la covariance de l'échantillon pour deux variables quelconques dépend des liens des deux variables en question avec les variables du groupement  $z_i$ . On devrait s'attendre à ce que les effets cumulatifs soient plus importants quand elles sont plus étroitement liées aux variables de groupement. En raison de la condition  $S_{zz}^z \approx \Sigma_{zz}^z$  les variables 2 échappent aux effets de la sélection et de l'agrégation. Néanmoins, il est peu probable que ces conditions s'appliquent dans la réalité, de sorte que beaucoup de variables auront un biais. L'erreur d'échantillonnage et de groupement attribuable aux variables auxiliaires est mesurée par  $S_{zz}^z - \Sigma_{zz}^z$  pour l'estimateur de niveau unitaire, et par  $S_{zz}^z - \Sigma_{zz}^z$  pour l'estimateur au niveau du groupe. L'expression  $S_{zz}^z - \Sigma_{zz}^z$  traduit l'effet net de l'échantillonnage et de l'agrégation sur les variables auxiliaires.

Le deuxième terme de biais (2.36) est nul si  $\Delta_{yz,z} = 0$ . Il n'existe donc aucune corrélation résiduelle à l'intérieur du groupe pour les variables  $y$ , sous réserve des variables de groupement. La chose est peu probable en pratique, mais il est souhaitable d'identifier les variables de groupement qui intègrent le plus possible les effets cumulatifs en rendant le terme résiduel le plus faible possible.

Les effets du groupement et de l'échantillonnage selon  $z$  et l'effet attribuable à la corrélation résiduelle à l'intérieur du groupe s'additionnent; c'est ce qui se produit avec les types de corrélations intérieures au groupe plus complexes, pourvu que la linéarité du modèle soit préservée. Si  $z$  suit un simple modèle des composantes de la variable comme le modèle A,

$$E[S_{zz}^z | s, c] = \Sigma_{zz}^z + (\pi^* - 1)\Delta_{zz}^z$$

$$E[S_{yy}^z | s, c] = \Sigma_{yy}^z + (\pi^* - 1)\beta_{yz}'\Delta_{zz}^z\beta_{yz} + \Delta_{yy,z}^z$$

(2.37)

et les covariances des variables intéressantes à l'intérieur du groupe seront constituées de l'élément résultant des covariances des variables auxiliaires à l'intérieur du groupe et des composantes résiduelles. Le côté droit de l'équation (2.37) est tiré de (2.18) puisque  $z$  suit sans conditions le modèle des composantes de la variance, si  $y$  en fait autant. Le modèle fondamental a pour but de trouver les variables auxiliaires qui permettront de réduire au maximum les covariances résiduelles ou conditionnelles  $\Delta_{yy,z}^z$  à l'intérieur d'un groupe, ou de les éliminer, dans le meilleur des cas.

## 2.5 Correction des effets cumulatifs

Peu de propositions utiles ont été avancées sur la manière dont les analyses de niveau régional pourraient être corrigées afin de donner une estimation raisonnable

des relations au niveau unitaire. Duncan et Davis (1953) ont examiné la fourchette éventuelle des coefficients de corrélation au moyen d'un tableau carré  $(2 \times 2)$  dont les marges étaient connues. Les limites résultantes sont souvent trop importantes pour présenter une utilité pratique. De son côté, Goodman (1959) a énoncé les conditions précises dans lesquelles l'analyse écologique pourrait servir à tirer des inférences sur les relations de niveau individuel avec un modèle de régression. Langbein et Litchman (1978) ont examiné quelques méthodes applicables aux groupements articulés sur les variables dépendantes, quand on connaît les variances de niveau unitaire des variables dépendantes et indépendantes du modèle de régression. Malheureusement, aucune de ces approches n'est assez générale pour permettre la résolution du problème.

Si on examine le biais pour  $S_{yy}^z$  dans (2.36), on constate qu'en ajoutant  $\beta_{yz}'(\Sigma_{zz}^z - S_{zz}^z)\beta_{yz}$  à  $S_{yy}^z$ , le terme de biais attribuable aux variables de groupement disparaît. L'équation (2.31) implique que

$$E[\bar{B}_{yz} | s, z, c] = \beta_{yz} \quad (2.38)$$

où  $\bar{B}_{yz} = S_{zz}^{-1} S_{zy}^z$ .

Si on disposait de la matrice des covariances  $S_{zz}^{zs_0}$  de  $z$ , pour l'échantillon unitaire  $s_0$  tiré de  $m_0$  groupes, l'estimateur corrigé

$$\bar{\Sigma}_{yy}^z(z) = \bar{S}_{yy}^z + \bar{B}_{yz}'(S_{zz}^{zs_0} - \bar{S}_{zz}^z)\bar{B}_{yz} \quad (2.39)$$

éliminerait le biais d'agrégation résultant des variables de groupement  $z$ , pourvu que  $S_{zz}^{zs_0}$  s'approche de  $\Sigma_{zz}^z$ . Il se peut que la source de  $S_{zz}^{zs_0}$  soit largement indépendante des données utilisées dans  $S_{yy}^z$  et  $\bar{B}_{yz}$ . Steel (1985) montre que l'estimateur corrigé (2.39) peut correspondre à l'estimateur du maximum de vraisemblance de  $\Sigma_{yy}^z$  (après substitution de  $m - 1$  par  $m$ , etc., comme on le fait d'habitude). Si la normalité de la répartition de  $(y, z)$  s'applique,  $s_0$  devient un simple échantillon aléatoire de la population et  $\Delta_{yy,z}^z = 0$ . L'estimateur corrigé correspond à l'ajustement de Pearson (1903) que Holt, Smith et Winter (1980) envisagent dans leur analyse de régression et que Smith et Holmes (1989) utilisent dans leur analyse à plusieurs variables. Dans les deux cas, on corrige les statistiques obtenues à partir des données de niveau unitaire issues d'un échantillon prélevé en fonction des variables auxiliaires. En ce qui nous concerne, la correction est appliquée aux statistiques tirées des moyennes des régions, et les variables auxiliaires utilisées pour effectuer la correction comprennent les variables de groupement ainsi que les variables du plan d'échantillonnage. L'estimateur corrigé de  $\mu_y$  est

$$\mu_y(z) = \bar{y} + \bar{B}_{yz}'(\bar{z}_{s_0} - z) \quad (2.40)$$

où  $\bar{z}_{s_0}$  est la moyenne de  $s_0$ .

D'après (2.34) et (2.38), on constate que

$$E[\mu_y(z) | s, z, s_0, c] = \mu_y + \beta_{yz}'(\bar{z}_{s_0} - \mu_z). \quad (2.41)$$



formation du groupe se caractérise par les variables auxiliaires  $z_i$ . Les variables auxiliaires peuvent être considérées comme des variables déterminant à quel groupe appartient telle ou telle unité. Sous un angle plus général, les variables auxiliaires correspondent aux principales variables de niveau individuel qui ne sont pas distribuées au hasard ni entre les groupes à cause des processus de sélection ou de migration auxquels la population est assujettie. On peut aussi inclure au modèle les variables contextuelles sous la forme de variables auxiliaires de valeur identique pour toutes les unités du groupe.

Si le vecteur des variables auxiliaires a une distribution marginale dont la moyenne est  $\mu_z$  et la matrice des covariances,  $\Sigma_{zz}$  la moyenne marginale et la matrice des covariances de  $y$  seront respectivement  $\mu_y = \mu_{y,z} + \beta'_{yz}\mu_z$  et  $\Sigma_{yy} = \Sigma_{y,z} + \beta'_{yz}\Sigma_{zz}\beta_{yz}$ . Les propriétés des moyennes de l'échantillon au niveau du groupe sont faciles à tirer du modèle B :

$$E[\bar{y}_g | s, z, c] = \mu_y + \beta'_{yz}(\bar{z}_g - \mu_z) \quad (2.22)$$

$$V(\bar{y}_g | s, z, c) = \frac{1}{n_g} \Sigma_{yy,z} \quad (2.23)$$

$$\text{Cov}(\bar{y}_g, \bar{y}_h | s, z, c) = 0 \quad g \neq h. \quad (2.24)$$

Les statistiques au niveau du groupe auront donc les propriétés qui suivent :

$$E[\bar{y} | s, z, c] = \mu_y + \beta'_{yz}(\bar{z} - \mu_z) \quad (2.25)$$

$$E[S_{yy} | s, z, c] = \Sigma_{yy} + \beta'_{yz}(S_{zz} - \Sigma_{zz})\beta_{yz} \quad (2.26)$$

$$E[\bar{S}_{yy} | s, z, c] = \Sigma_{yy} + \beta'_{yz}(\bar{S}_{zz} - \Sigma_{zz})\beta_{yz} \quad (2.27)$$

où  $S_{zz}$  et  $\bar{S}_{zz}$  sont définis de façon analogue à  $S_{yy}$  et  $\bar{S}_{yy}$  de l'équation (2.3) et de la phrase suivant cette dernière.

## 2.4 Modèle combiné

Jusqu'à présent, on peut dire que les deux modèles examinés expliquent les effets de groupe chacun de leur côté. Il est possible de les combiner en un modèle plus réaliste qui intégrera les deux effets de groupe et les composantes résiduelles de la variance :

### Modèle C :

$$E[y_i | z, c] = \mu_{y,z} + \beta'_{yz} z_i \quad (2.28)$$

$$V(y_i | z, c) = \Sigma_{yy,z} \quad (2.29)$$

$$\text{Cov}(y_i, y_j | z, c) = \Delta_{yy,z} \quad \text{si } c_i = c_j \quad i \neq j \quad (2.30)$$

$$= 0 \quad \text{sinon.}$$

Le nouveau modèle accepte les mécanismes de création des groupes que caractérisent les variables auxiliaires  $z_i$ . On y retrouve aussi les corrélations résiduelles à l'intérieur d'un groupe qui résultent des effets aléatoires attribués aux variables aléatoires inconnues du niveau du groupe, après exclusion des variables de groupement.

Voici les propriétés des moyennes de l'échantillon au niveau du groupe provenant du modèle C, lorsqu'on néglige l'échantillonnage, étant donné  $(z, c)$ ,

$$E[\bar{y}_g | s, z, c] = \mu_y + \beta'_{yz}(\bar{z}_g - \mu_z) \quad (2.31)$$

$$V(\bar{y}_g | s, z, c) = \frac{1}{n_g} (\Sigma_{yy,z} + (n_g - 1)\Delta_{yy,z}) \quad (2.32)$$

$$\text{Cov}(\bar{y}_g, \bar{y}_h | s, z, c) = 0 \quad g \neq h \quad (2.33)$$

$$E[\bar{y} | s, z, c] = \mu_y + \beta'_{yz}(\bar{z} - \mu_z) \quad (2.34)$$

$$E[S_{yy} | s, z, c] = \Sigma_{yy} + \beta'_{yz}(S_{zz} - \Sigma_{zz})\beta_{yz} \quad (2.35)$$

$$- \frac{n^0 - 1}{n - 1} \Delta_{yy,z} \quad (2.36)$$

$$E[\bar{S}_{yy} | s, z, c] = \Sigma_{yy} + \beta'_{yz}(\bar{S}_{zz} - \Sigma_{zz})\beta_{yz} + (n^* - 1)\Delta_{yy,z}. \quad (2.36)$$

Les équations (2.17) et (2.18) montrent comment l'agré-gation amplifie les effets aléatoires au niveau du groupe dans le modèle des composantes de la variance A. Dans l'équation (2.17), le coefficient de  $\Delta_{yy}$  est  $0(m^{-1})$ , tandis que dans (2.18), il correspond à  $0(n)$ . Avec le modèle combiné C, les équations (2.35) et (2.36) montrent comment l'inclusion des variables de groupement permet de scinder le biais en deux éléments additifs : le premier associé aux variables de groupement, aux liens de ces dernières avec les variables d'intérêt et à leur effet cumulatif, et le second faisant intervenir  $\Delta_{yy,z}$ , soit les composantes résiduelles de la variance, après prise en compte des variables de groupement. Notons que les coefficients de  $\Delta_{yy,z}$  dans les équations (2.35) et (2.36) restent  $0(m^{-1})$  et  $0(n)$  respectivement, comme aux équations (2.17) et (2.18). Les composantes résiduelles de la variance devraient néanmoins être plus faibles, en général. L'équation (2.29) part de l'hypothèse que la variance résiduelle de  $c$  est constante. L'hypothèse que l'échantillonnage est négligeable pour  $(z, c)$  signifie que le plan d'échantillonnage peut dépendre des variables auxiliaires et des paramètres discriminants du groupe. On peut donc effectuer une stratification selon  $z$  et procéder à un échantillonnage en grappes ou à plusieurs degrés en fonction des groupes.

Les propriétés des statistiques au niveau unitaire et au niveau du groupe sont les suivantes:

$$\text{Cov}(\bar{y}_g, \bar{y}_h | s, c) = 0 \quad g \neq h. \quad (2.15)$$

$$E[\bar{y} | s, c] = \mu_y \quad (2.16)$$

$$E[S_{yy} | s, c] = \Sigma_{yy} - \frac{n-1}{n^0 - 1} \Delta_{yy} \quad (2.17)$$

$$E[S_{yy} | s, c] = \Sigma_{yy} + (n^* - 1) \Delta_{yy} \quad (2.18)$$

où  $n = n/m$ ,  $n^0 = 1/n \sum_{g \in s} n_g^2 = n(1 + C_n^2)$ ,  $n^* = n(1 - C_n^2)/(m - 1)$  et  $C_n^2 = 1/m \sum_{g \in s} (n_g - n)^2/n^2$  est le carré du coefficient de variation des tailles de groupe de l'échantillon. Notons que le coefficient de  $\Delta_{yy}$  est  $0(m^{-1})$  dans (2.17) et  $0(n)$  dans (2.18), ce qui montre bien comment une petite erreur d'analyse au niveau unitaire prend de l'ampleur après aggrégation. Nous approfondirons ces résultats à la partie 2.4.

### 2.3 Modèles de groupement

Les chercheurs qui s'intéressent à l'analyse écologique ont échafaudé des modèles qui tiennent compte du mécanisme de création des groupes. Ils supposent qu'un processus de synthèse attribue les unités à tel ou tel groupe, selon un vecteur de variables de groupement  $z_i$ , de façon stochastique ou déterministe. Blalock (1964) recourt implicitement à cette approche dans son analyse, tandis qu'Hannan et Bursstein (1974), Litchman (1974), Langbein et Litchman (1978), Smith (1977) et Blalock (1979, 1985) s'en servent de façon explicite. Steel (1985) parle de modèles de groupement, car il suppose que les groupes naissent au terme d'un processus quelconque qui fait intervenir les variables au niveau des relations à l'étude. La formation du groupe est perçue comme une distorsion, aussi les relations intéressantes sont-elles définies auparavant. On précise souvent dans la discussion sur les modèles contextuels que les effets contextuels apparents peuvent résulter de tels facteurs. La version multivariée du modèle est la suivante:

#### Modèle B:

$$E[y_i | z, c] = \mu_{y,z} + \beta'_{yz} z_i \quad (2.19)$$

$$V(y_i | z, c) = \Sigma_{yy,z} \quad (2.20)$$

$$\text{Cov}(y_i, y_j | z, c) = 0 \quad i \neq j. \quad (2.21)$$

Dans ce modèle, l'espérance conditionnelle de  $y_i$  dépend seulement de la valeur des variables auxiliaires de la  $i$ -ième unité et est indépendante du groupe à laquelle l'unité appartient ou de la valeur des variables auxiliaires des autres unités de la population. La covariance conditionnelle entre deux unités quelconques est zéro. Ce modèle est valable pour les modèles de groupement dans lesquels le mécanisme de

Ces propriétés s'appliquent si on peut négliger l'échantillonnage étant donné les paramètres discriminants du groupe, bref le plan d'échantillonnage peut dépendre des groupes mais pas de  $y$  ou d'une variable quelconque associée à  $y$ , sous réserve de  $c$ . Par exemple, on pourrait utiliser le recensement ou un échantillon aléatoire simple de groupes et d'unités appartenant à ces groupes. On peut recourir à des statistiques non pondérées au niveau du groupe en posant  $n_g^* = 1$  dans les équations (2.2) et (2.3). On obtient ainsi des estimateurs inefficaces. Le degré d'inefficacité dépendra de la distribution des groupes d'échantillon des groupes. La pondération selon les tailles d'échantillon des groupes est importante. Cela fait, on peut procéder aux inférences habituelles en apportant les corrections appropriées aux degrés de liberté. La variabilité dépend du nombre de régions plutôt que du nombre d'observations et on adapte les intervalles de confiance et les épreuves en conséquence.

### 2.2 Modèle des composantes de la variance

Une façon simple d'illustrer la corrélation positive entre les membres d'un groupe normalement observée dans une population consiste à utiliser un modèle des composantes de la variance, ce qui, dans le cas d'une analyse à plusieurs variables, correspond à

$$y_i = \mu_y + v_g + \epsilon_i \quad i \in g$$

où  $v_g$  et  $\epsilon_i$  sont des composantes aléatoires indépendantes au niveau du groupe et au niveau de l'individu, respectivement, avec une espérance mathématique nulle,  $V(\epsilon_i | c) = \Sigma_{\epsilon\epsilon}$  et  $V(v_g | c) = \Delta_{yy}$ .

#### Modèle A:

$$E[y_i | c] = \mu_y \quad (2.10)$$

$$V(y_i | c) = \Sigma_{\epsilon\epsilon} + \Delta_{yy} = \Sigma_{yy} \quad (2.11)$$

$$\text{Cov}(y_i, y_j | c) = \Delta_{yy} \quad \text{si } c_i = c_j \quad i \neq j \quad (2.12)$$

$$= 0 \quad \text{sinon.}$$

La notation  $V(\cdot | c)$  signifie que la matrice des covariances dépend de l'étiquette  $c$  du groupe, donc détermine l'appartenance à un groupe commun. On estime cependant qu'elle est inconditionnelle pour les effets aléatoires au niveau du groupe. Par conséquent,  $V(y_i | c)$  inclut la variance totale, à la fois pour la matrice des covariances de groupe  $\Sigma_{\epsilon\epsilon}$  et la matrice des covariances de groupe  $\Delta_{yy}$ . On obtient facilement les propriétés des moyennes de l'échantillon au niveau du groupe à partir du modèle A si on peut négliger l'échantillonnage étant donné  $c$ ,

$$E[\bar{y}_g | s, c] = \mu_y \quad (2.13)$$

$$V(\bar{y}_g | s, c) = \frac{1}{n_g} (\Sigma_{yy} + (n_g - 1) \Delta_{yy}) \quad (2.14)$$



Les premiers travaux sur l'analyse des données agrégées remontent à Gehlke et Biehl (1934). D'éminents chercheurs

(1964), Openshaw et Taylor (1979) et, plus récemment, Arbia (1989) ont sensiblement contribué à l'étude du problème. Le fait que les unités régionales utilisées présentent rarement une importance particulière (elles sont construites pour des raisons d'économie, de facilité ou de commodité administrative) soulève aussi des difficultés. Par ailleurs, les résultats de l'analyse effectuée au niveau du groupe dépendent de l'échelle des unités, soit de leur taille moyenne et du jeu de limites retenu. Plusieurs études empiriques ont fait ressortir ces aspects, notamment celles de Clark et Avery (1976), de Perle (1977), d'Openshaw (1984) et de Fortheringham et Wong (1991). Les travaux n'ont malheureusement pas débouché sur une théorie permettant une application générale, ni sur des méthodes pratiques en vertu desquelles on pourrait tirer des inférences fiables de niveau unitaire à partir des résultats des analyses effectuées au niveau du groupe.

On attribue les effets de l'aggrégation au fait que la population des unités géographiques ne s'est pas constituée au hasard. En règle générale, les individus d'une même région ont tendance à se ressembler davantage, car ils n'ont pas choisi de vivre là au hasard ou parce qu'ils s'y trouvent consécutivement à l'action de forces analogues, des différences socioéconomiques entre les régions, différences qui se confondent avec les effets individuels dans l'analyse statistique des données agrégées sur les régions concernées. Nous proposons d'abord un modèle général simple qui s'efforce d'intégrer ces effets, puis nous examinons les conséquences de son utilisation et les implications d'une telle méthode pour l'analyse au niveau régional. Par ailleurs, nous suggérons des méthodes qui, dans certaines circonstances, donneront une estimation non biaisée des paramètres de niveau individuel à partir des données agrégées, donc permettront d'éviter l'erreur écologique. Ces méthodes font intervenir des variables auxiliaires pour lesquelles certaines sources donnent une matrice des covariances de niveau unitaire pour l'échantillon. Cette approche a été appliquée aux données du recensement de 1991 du Royaume-Uni et une stratégie a été développée en vue de l'analyse des données agrégées.

## 2. MODÈLES POUR LES EFFETS RÉGIONAUX

Soit une population de  $N$  sujets ayant chacun un vecteur  $y$  de caractéristiques à étudier. La population se compose de  $M$  groupes et la variable aléatoire  $c_i$  indique la région à laquelle la  $i$ -ème unité de population appartient. On dénombre  $N_g$  individus dans la  $g$ -ième région. Supposons que  $\mu_y$  et  $\Sigma_{yy}$  soient les paramètres de la superpopulation. Compte tenu de ces prémisses, on obtient la théorie statistique qui suit. Quelques aspects relatifs au plan d'échantillonnage seront toutefois examinés à la fin de la partie 2.

moyenne de la  $g$ -ième région:

$$\bar{y}_g = \frac{1}{n_g} \sum_{i \in g} y_i \quad (2.1)$$

moyenne totale de l'échantillon:

$$\bar{y} = \frac{1}{n} \sum_{i \in g} n_g \bar{y}_g = \frac{1}{n} \sum_{i \in g} y_i \quad (2.2)$$

matrice des covariances de l'échantillon au niveau régional:

$$S_{yy} = \frac{1}{I} \sum_{g \in g} n_g (\bar{y}_g - \bar{y})(\bar{y}_g - \bar{y})' \quad (2.3)$$

On peut définir des statistiques analogues au niveau unitaire, mais l'analyste ne pourra s'en servir. Par exemple,  $S_{yy} = 1/(n-1) \sum_{i \in g} (y_i - \bar{y})(y_i - \bar{y})'$  est la matrice des covariances de l'échantillon au niveau unitaire.

### 2.1 Groupement aléatoire

Bien que les groupes géographiques soient rarement le fait du hasard, pareille hypothèse constitue un bon point de départ quand on s'intéresse à l'analyse écologique. Si les groupes se forment de façon aléatoire, maintes analyses effectuées au niveau du groupe seront valables, même si leur utilité est réduite. Steel et Holt (1995) étudient les propriétés de diverses statistiques comme la moyenne, la variance et les coefficients de régression et de corrélation dans une situation de ce genre. Quand les groupes se constituent au hasard, c'est-à-dire  $y \perp c$ , alors

$$E[\bar{y}_g | s, c] = \mu_y \quad (2.4)$$

$$V(\bar{y}_g | s, c) = \frac{1}{I} \Sigma_{yy} \cdot n_g \quad (2.5)$$

Les propriétés fondamentales des statistiques au niveau unitaire et au niveau du groupe sont faciles à déduire:

$$\text{Cov}(\bar{y}_g, \bar{y}_h | s, c) = 0 \quad g \neq h \quad (2.6)$$

$$E[\bar{y} | s, c] = \mu_y \quad (2.7)$$

$$E[S_{yy} | s, c] = \Sigma_{yy} \quad (2.8)$$

$$E[S_{yy} | s, c] = \Sigma_{yy} \cdot \quad (2.9)$$

# Inférences au niveau unitaire à partir de données agrégées

D.G. STEEL, D. HOLT et M. TRAMMER<sup>1</sup>

## RÉSUMÉ

Les données ne sont souvent disponibles que sous la forme de moyennes de groupes ou de régions. Or, on sait pertinemment qu'une analyse statistique articulée sur les données de ce genre aboutit fréquemment à des résultats très différents de ceux obtenus lorsqu'on analyse les données correspondantes sur les individus ou les ménages. Croire que les résultats d'une analyse de niveau régional sont applicables au niveau individuel, c'est risquer de commettre l'erreur écologique. Les effets de l'aggrégation ou les effets écologiques résultent en partie du fait qu'une région n'est pas constituée d'un assemblage aléatoire d'êtres humains ou de ménages. Les paramètres socioéconomiques varient considérablement d'une région à l'autre. On doit intégrer la structure de la population au modèle statistique utilisé pour l'analyse si on veut bien saisir les conséquences de l'aggrégation. Les auteurs proposent un modèle général simple pour y parvenir et décrivent l'effet de ce modèle sur l'estimation des moyennes et des matrices des covariances de la population. Par ailleurs, ils montrent comment obtenir une estimation non biaisée des paramètres au niveau de l'individu à partir des données agrégées, de façon à éviter l'erreur écologique. Les méthodes qu'ils préconisent supposent l'identification des «variables de groupement» qui caractérisent le processus qui a mené à la structure de la population ou, du moins, les différences entre régions. On doit pour cela trouver une estimation de la matrice des covariances pour les variables d'aggrégation, au niveau unitaire, d'une source quelconque. L'analyse des données du recensement de 1991 du Royaume-Uni a permis d'identifier les principales variables d'aggrégation et de mesurer l'efficacité des méthodes de correction envisagées pour estimer les matrices des covariances et les coefficients de corrélation. Les résultats de ces travaux concourent à l'élaboration d'une stratégie pour l'analyse des données agrégées.

MOTS CLÉS: Aggrégation; erreur écologique; groupement; sélection; composantes de la variance.

## 1. INTRODUCTION

Les chercheurs qui souhaitent étudier les relations au niveau de l'individu éprouvent souvent des difficultés parce qu'ils doivent utiliser des données agrégées, par exemple une moyenne ou un total régional. Idéalement, on devrait se servir des données au niveau de l'unité recueillies lors du sondage ou du recensement, mais ces données sont inaccessibles parce qu'il faut en préserver la confidentialité ou parce qu'elles ne viennent pas d'une enquête ou d'un recensement récents. Les régimes administratifs nous procurent des renseignements sur diverses variables, par exemple le chômage, la santé et la morbidité. Malheureusement, ces données sont fournies habituellement pour des agrégats, notamment par région, toujours pour des raisons de confidentialité. Le recensement fournit lui aussi des données régionales. C'est pourquoi les recherches sociales et épidémiologiques portent encore couramment sur des données collectives.

Soit une population au sein de laquelle chaque sujet est associé à un vecteur de variables intéressantes, dont la répartition a pour moyenne  $\mu_y$  et pour matrice des covariances  $\Sigma_{yy}$ . Nous aimerions savoir quels liens existent entre ces variables, comme les représentent les corrélations, les coefficients de régression et les grandes composantes qu'on peut tirer de la matrice des covariances  $\Sigma_{yy}$ , sujet principal de nos inférences. Ces variables pourraient comprendre, par exemple, une batterie de tests d'accomplissement dans le cadre d'une étude sur l'éducation,

L'incidence d'une maladie donnée et une suite de variables explicatives dans le contexte d'une étude épidémiologique, voire une série de mesures de privation dans l'optique d'une étude sociologique. Nous supposons ici qu'on ne peut obtenir les données au niveau de l'individu. Néanmoins, la région peut être subdivisée en plus petits ensembles comme des districts de recensement (DR) et pour chacun de ces sous-ensembles  $g$  ou pour un échantillon de régions, on peut calculer le vecteur des valeurs moyennes  $y_g$  pour les variables étudiées de l'échantillon  $n_g$  d'où sont extraites ces dernières.

Le but de l'analyse est d'obtenir une matrice des covariances  $\Sigma_{yy}$  couvrant ces petites régions. L'inférence n'est pas conditionnelle à l'appartenance à une petite région, mais se rapporte à la distribution marginale entre ces régions. La situation est différente avec l'estimation d'une petite région. Dans ce cas, l'inférence vise la distribution conditionnelle dans la région en question. Pareil objectif est tout à fait légitime, mais diffère de celui qui nous intéresse. Il se peut qu'on puisse recourir aux mêmes modèles, mais l'inférence n'a pas la même visée. Notre formulation nous permet néanmoins d'ajouter les variables spécifiques à certains groupes aux variables intéressantes, s'il y a lieu. Ainsi, si on associe à chaque individu un ensemble de moyennes DR pour la région où il se trouve, il est alors possible de les inclure au vecteur  $y$  qui nous intéresse. Pareille approche se prête aux analyses de régression qui utilisent les moyennes des petites régions comme variables explicatives.

<sup>1</sup> D.G. Steel, Department of Applied Statistics, University of Wollongong, NSW 2522, Australia; D. Holt et M. Trammer, Department of Social Statistics, University of Southampton, SO17 1BJ, United Kingdom.



Dans un article portant sur la stratification optimale, Sianta et Krenzke discutent de l'utilisation de la méthode Lavalée-Hidroglo. Cette méthode itérative vise à déterminer des bornes de stratification qui minimisent la taille de l'échantillon pour un coefficient de variation donné. Dans un contexte pratique, les auteurs présentent les difficultés rencontrées avec la méthode et montrent comment elles ont été résolues.

Dagum propose une nouvelle méthode pour estimer les tendances sous-jacentes à partir des données désaisonnalisées. Son approche comporte deux étapes. On extrapole d'abord les données désaisonnalisées au moyen d'un modèle ARMMI. Ensuite, on applique un filtre Henderson à 13 termes à la série élargie en se servant de strictes limites sigma pour identifier et remplacer les valeurs extrêmes. L'auteur compare sa méthode à la méthode type au moyen de données issues de plusieurs séries économiques chronologiques et constate ainsi que la nouvelle méthode engendre moins d'irrégularités dans la tendance estimative tout en permettant une identification aussi rapide des changements de sens et en exigeant généralement de plus petites corrections.

Tillé propose un algorithme généralisant la méthode de sélection-rejet permettant de constituer un échantillon aléatoire simple sans remise. Un cas particulier de cet algorithme baptisé «algorithme de stratification mobile» est discuté. Il permet d'obtenir un effet de stratification lissé en utilisant comme variable de stratification le numéro d'ordre des unités d'observation. Cet algorithme permet de contourner le délicat problème du découpage d'une variable continue en strates.

De Waal et Wilkenborg passent en revue les recherches récentes sur le contrôle statistique de la divulgation dans les fichiers de microdonnées, selon le point de vue du bureau de statistique néerlandais. Les auteurs développent des modèles en fonction de la probabilité qu'on puisse identifier un enregistrement particulier et de la probabilité qu'un enregistrement quelconque dans un fichier de microdonnées puisse être identifié de nouveau. Un nouveau codage global et la suppression locale sont des méthodes qui permettent de réduire les risques de divulgation. Les auteurs parviennent à la conclusion que beaucoup de recherche et de travaux de développement restent à faire sur le plan de la méthodologie avant qu'on parvienne à des logiciels efficaces.

Enfin, c'est le cœur lourd que je dois signaler le récent décès de Maria Gonzalez, emportée par un infarctus alors qu'elle passait ses vacances à Porto Rico, en février dernier. Parmi ses nombreuses contributions à la statistique, ces dernières années, Maria était notamment devenue rédactrice associée de *Techniques d'enquête*. À ce titre, nous avons considérablement apprécié les efforts qu'elle a déployés pour améliorer la qualité et le champ d'intérêt du périodique. Nous la regretterons longtemps. Sa notice nécrologique, rédigée par Elizabeth et Fritz Scheuren, a paru dans le numéro d'avril de *Amslat News*.

Le rédacteur en chef

## Dans ce numéro

Ce numéro de *Techniques d'enquête* contient des articles traitant de divers sujets. Dans le premier, Steel, Holt et Tramter examinent le problème de l'utilisation de données agrégées lors d'études portant sur les relations au niveau de l'individu ou du ménage. Ils proposent un modèle général simple qui s'efforce d'intégrer les effets géographiques d'agrégation. Ils décrivent ensuite l'effet de ce modèle sur l'estimation des moyennes et des matrices de covariances de la population, ainsi que sur l'analyse au niveau régional. De plus, en faisant intervenir des variables auxiliaires pour lesquelles certaines sources externes fournissent une estimation de la matrice de covariances au niveau unitaire, les auteurs proposent des méthodes qui donnent une estimation non biaisée des paramètres de niveau individuel, de façon à éviter l'effet d'agrégation géographique.

Bindar propose une approche de type «recette» permettant d'obtenir une approximation, par série de Taylor, de la variance d'une grande variété d'estimateurs extraits d'enquêtes complexes. L'auteur présente plusieurs exemples utiles et de nouveaux résultats sur l'application de cette technique générale à l'échantillonnage à deux phases. Il justifie aussi la méthode proposée en montrant qu'elle est cohérente avec la formulation avancée antérieurement par Bindar et Patak.

De leur côté, Yung et Rao proposent une approximation linéaire de l'estimateur de variance jackknife. La méthode du jackknife linéarisée garde les propriétés statistiques intéressantes de l'estimateur de variance jackknife habituel, mais nécessite des calculs beaucoup moins laborieux. Le nouvel estimateur de la variance que proposent les auteurs s'applique à un estimateur de régression généralisée d'un total et au ratio de deux estimateurs de régression généralisée. Dans le cadre d'une simulation s'inspirant des données de l'enquête sur la U.S. Current Population Survey, ces auteurs ont constaté que l'estimateur de variance jackknife, l'estimateur de variance jackknife linéarisée et l'estimateur de variance linéarisée habituel donnent de très bons résultats pour l'estimation d'un total obtenu par stratification a posteriori, tandis qu'une forme incorrecte de l'estimateur jackknife entraîne un biais important.

Chaubey, Nebebe et Chen envisagent l'utilisation d'un modèle gaussien inverse pour les données étalées vers la droite et élaborent un estimateur correspondant au moyen d'un modèle pour les totaux de domaines, estimateur qui se compose de variables explicatives pour une régression gaussienne inverse et d'estimateurs d'expansion du biais dû à la régression. Les auteurs proposent aussi une variante de l'estimateur qui attribue un poids réduit aux paramètres servant à corriger le biais. Cet estimateur ressemble à l'estimateur de régression modifié que proposaient Särndal et Hidiroglou. Les estimateurs suggérés fonctionnent raisonnablement bien dans le cadre d'une simulation reposant sur des données synthétiques relatives au revenu qui s'inspire de l'enquête de Statistique Canada sur le revenu, l'équipement ménager et les finances des ménages.

Rizzo, Kalton et Brick étudient l'usage des renseignements complémentaires pour compenser la non-réponse du panel par diverses techniques d'ajustement des poids. Se servant de données tirées du Survey of Income and Program Participation (SIPP) pour illustrer leur méthode de pondération, les auteurs abordent deux aspects importants, à savoir le choix des variables auxiliaires à utiliser lors de l'ajustement des poids servant à compenser la non-réponse et le choix de la technique d'ajustement combinée à un modèle de régression logistique. Les méthodes d'ajustement de la pondération des non-réponses examinées s'appuient sur des modèles de régression logistique, des algorithmes de recherche catégorique et une itération généralisée. Les auteurs analysent en détail la comparaison empirique des diverses méthodes.

Ding et Fienberg examinent les modèles de l'erreur d'appariement dont on peut se servir pour estimer la population totale à partir de l'appariement probabiliste de deux échantillons ou plus. Leur modèle est principalement destiné à être appliqué à un recensement par échantillonnage multiple, c'est-à-dire à un recensement double d'échantillons secondaires. Ils illustrent l'utilité de leurs méthodes en les appliquant à une analyse des données issues de la répétition générale du recensement de 1988 à St. Louis, pour laquelle trois échantillons ont été appariés: celui du recensement proprement dit, l'échantillon de l'enquête postcensitaire et la liste administrative complémentaire.





# TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada  
Volume 22, numéro 1, juin 1996

## TABLE DES MATIÈRES

Dans ce numéro .....	1
D.G. STEEL, D. HOLT et M. TRAMMER	
Inférences au niveau unitaire à partir de données agrégées .....	3
D.A. BINDER	
Méthodes de linéarisation pour les échantillons à une et deux phases: Une approche de type «recette» .....	17
W. YUNG et J.N.K. RAO	
Linéarisation des estimateurs de variance jackknife dans un échantillonnage stratifié à degrés multiples .....	23
Y.P. CHAUBEY, F. NEBBE et P.S. CHEN	
Estimation des caractéristiques des petites régions au moyen d'un modèle gaussien inverse .....	33
L. RIZZO, G. KALTON et J.M. BRICK	
Comparaison de quelques méthodes de correction de la non-réponse d'un panel .....	43
Y. DING et S.E. FIENBERG	
Estimation de la population par échantillonnage multiple et sous-dénombrement lors du recensement en présence d'erreurs d'appariement .....	55
J.G. SLANTA et T.R. KRENZKE	
Utilisation de la méthode de Lavallée et Hidiroglou pour le calcul des limites de stratification aux fins de l'enquête annuelle sur les dépenses en capital du Bureau of the Census .....	65
E.B. DAGUM	
Nouvelle méthode visant à limiter le nombre d'ondulations indésirables et les corrections lors de l'estimation de la tendance-cycle au moyen du modèle X-11-ARMMI .....	77
Y. TILLÉ	
Un algorithme de stratification mobile .....	85
A.G. de WAAL et L.C.R.J. WILLENBORG	
Apérçu du contrôle statistique de la divulgation des microdonnées .....	95



# TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Techniques d'enquête est répertoriée dans The Survey Statistician et Statistical Theory and Methods Abstracts. On peut en trouver les références dans Current Index to Statistics, et Journal Contents in Qualitative Methods.

## COMITÉ DE DIRECTION

Président

G.J. Brackstone

Membres

D. Binder

G.J.C. Hole

F. Mayda (Directeur de la Production)

D. Roy

R. Platek (Ancien président)

## COMITÉ DE RÉDACTION

Rédacteur en chef

M.P. Singh, Statistique Canada

Rédacteurs associés

D.R. Bellhouse, University of Western Ontario

D. Binder, Statistique Canada

J.-C. Deville, INSEE

J.D. Drew, Statistique Canada

J.-J. Droesbeke, Université Libre de Bruxelles

W.A. Fuller, Iowa State University

M. Gonzalez, U.S. Office of Management and Budget

R.M. Groves, University of Maryland

M.A. Hidiroglou, Statistique Canada

D. Holt, Central Statistical Office, U.K.

G. Kalton, Westat, Inc.

A. Mason, East-West Center

D. Pfeffermann, Hebrew University

Rédacteurs adjoints

J. Denis, M. Latouche, H. Mantel et D. Stukel, Statistique Canada

## POLITIQUE DE RÉDACTION

Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

## Présentation de textes pour la revue

Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à faire parvenir le texte rédigé en anglais ou en français au rédacteur en chef, M. M.P. Singh, Division des méthodes d'enquêtes-ménages, Statistique Canada, Tunney's Pasture, Ottawa (Ontario), Canada K1A 0T6. Prière d'envoyer quatre exemplaires dactylographiés selon les directives présentées dans la revue. Ces exemplaires ne seront pas retournés à l'auteur.

## Abonnement

Le prix de Techniques d'enquête (n° 12-001-XPB au catalogue) est de 45 \$ par année au Canada, 50 \$ (E.-U.) aux Etats-Unis, et de 55 \$ (E.-U.) par année à l'étranger. Prière de faire parvenir votre demande d'abonnement à Statistique Canada, Division des opérations et de l'intégration, Gestion de la circulation, 120, avenue Parkdale, Ottawa (Ontario), Canada K1A 0T6. Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête, l'American Association for Public Opinion Research et la Société Statistique du Canada.



Ottawa

ISSN 0714-0045

N° 12-001-XPB au catalogue  
Périodicité: semestrielle

Autres pays : 55 \$ US

États-Unis : 50 \$ US

Prix : Canada : 45 \$

Juin 1996

Tous droits réservés. Il est interdit de reproduire ou de transmettre le contenu de la présente publication, sous quelque forme ou par quelque moyen que ce soit, enregistrément sur support magnétique, reproduction électronique, mécanique, photographique, ou autre, ou de l'emmagasiner dans un système de recouvrement, sans l'autorisation écrite préalable des Services de concession des droits de licence, Division du marketing, Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

© Ministre de l'Industrie, 1996

Publication autorisée par le ministre  
responsable de Statistique Canada

JUN 1996 • VOLUME 22 • NUMÉRO 1

# UNE REVUE ÉDITÉE PAR STATISTIQUE CANADA

## TECHNIQUES D'ENQUÊTE









NUMÉRO 1

VOLUME 22

JUIN 1996

UNE REVUE  
ÉDITÉE  
PAR STATISTIQUE CANADA

Catégorie 12-001

---

# TECHNIQUES D'ENQUÊTE

---





12  
-001



# SURVEY METHODOLOGY

Catalogue 12-001-XPB

A JOURNAL  
PUBLISHED BY  
STATISTICS CANADA

DECEMBER 1996

•

VOLUME 22

•

NUMBER 2



Statistics  
Canada

Statistique  
Canada

Canada







---

# SURVEY METHODOLOGY

---

A JOURNAL  
PUBLISHED BY  
STATISTICS CANADA

DECEMBER 1996 • VOLUME 22 • NUMBER 2

Published by authority of the Minister  
responsible for Statistics Canada

© Minister of Industry, 1996

All rights reserved. No part of this publication may be reproduced,  
stored in a retrieval system or transmitted in any form or by any  
means, electronic, mechanical, photocopying, recording or otherwise  
without prior written permission from Licence Services,  
Marketing Division, Statistics Canada,  
Ottawa, Ontario, Canada K1A 0T6.

December 1996

Price: Canada: \$45.00  
United States: US\$50.00  
Other countries: US\$55.00

Catalogue no. 12-001-XPB  
Frequency: Semi-annual

ISSN 0714-0045

Ottawa



Statistics  
Canada

Statistique  
Canada

Canada

# SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is abstracted in The Survey Statistician and Statistical Theory and Methods Abstracts and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

## MANAGEMENT BOARD

**Chairman** G.J. Brackstone

**Members** D. Binder  
G.J.C. Hole  
F. Mayda (Production Manager)  
C. Patrick  
R. Platek (Past Chairman)  
D. Roy  
M.P. Singh

## EDITORIAL BOARD

**Editor** M.P. Singh, *Statistics Canada*

### Associate Editors

D.R. Bellhouse, <i>University of Western Ontario</i>	L.-P. Rivest, <i>Université Laval</i>
D. Binder, <i>Statistics Canada</i>	I. Sande, <i>Bell Communications Research, U.S.A.</i>
J.-C. Deville, <i>INSEE</i>	F.J. Scheuren, <i>George Washington University</i>
J.D. Drew, <i>Statistics Canada</i>	J. Sedransk, <i>Case Western Reserve University</i>
W.A. Fuller, <i>Iowa State University</i>	R. Sitter, <i>Simon Fraser University</i>
R.M. Groves, <i>University of Maryland</i>	C.J. Skinner, <i>University of Southampton</i>
M.A. Hidioglou, <i>Statistics Canada</i>	R. Valliant, <i>U.S. Bureau of Labor Statistics</i>
D. Holt, <i>Central Statistical Office, U.K.</i>	V.K. Verma, <i>University of Essex</i>
G. Kalton, <i>Westat, Inc.</i>	P.J. Waite, <i>U.S. Bureau of the Census</i>
R. Lachapelle, <i>Statistics Canada</i>	J. Waksberg, <i>Westat, Inc.</i>
S. Linacre, <i>Australian Bureau of Statistics</i>	K.M. Wolter, <i>National Opinion Research Center</i>
D. Pfeffermann, <i>Hebrew University</i>	A. Zaslavsky, <i>Harvard University</i>
J.N.K. Rao, <i>Carleton University</i>	

**Assistant Editors** J. Denis, M. Latouche, H. Mantel and D. Stukel, *Statistics Canada*

---

## EDITORIAL POLICY

Survey Methodology publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

## Submission of Manuscripts

Survey Methodology is published twice a year. Authors are invited to submit their manuscripts in either English or French to the Editor, Dr. M.P. Singh, Household Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Four nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

## Subscription Rates

The price of Survey Methodology (Catalogue no. 12-001-XPB) is \$45 per year in Canada, US \$50 in the United States, and US \$55 per year for other countries. Subscription order should be sent to Statistics Canada, Operations and Integration Division, Circulation Management, 120 Parkdale Avenue, Ottawa, Ontario, Canada K1A 0T6. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, and the Statistical Society of Canada.



**SURVEY METHODOLOGY**  
A Journal Published by Statistics Canada  
Volume 22, Number 2, December 1996

**CONTENTS**

In This Issue .....	105
<b>Weighting and Estimation</b>	
A.C. SINGH and C.A. MOHL Understanding Calibration Estimators in Survey Sampling .....	107
D.M. STUKEL, M.A. HIDIROGLOU and C.-E. SÄRNDAL Variance Estimation for Calibration Estimators: A Comparison of Jackknifing Versus Taylor Linearization .....	117
B.R. JAYASURIYA and R. VALLIANT An Application of Restricted Regression Estimation in a Household Survey .....	127
G. CHEN and J. CHEN A Transformation Method for Finite Population Sampling Calibrated With Empirical Likelihood .....	139
<hr/>	
K.J. THOMPSON and R. FISHER The Application of McNemar Tests to the Current Population Survey's Split Panel Study .....	147
J.L. ELTINGE and D.S. JANG Stability Measures for Variance Component Estimators Under a Stratified Multistage Design .....	157
Y.G. BERGER Asymptotic Variance for Sequential Sampling Without Replacement With Unequal Probabilities .....	167
A. COWLING, R. CHAMBERS, R. LINDSAY and B. PARAMESWARAN Applications of Spatial Smoothing to Survey Data .....	175
J.M. BRICK, J. WAKSBERG and S. KEETER Using Data on Interruptions in Telephone Service as Coverage Adjustments .....	185
G.S. PANDHER Optimal Sample Redesign Under GREG in Skewed Populations With Application .....	199
Acknowledgements .....	205





## In This Issue

This issue of *Survey Methodology* begins with a special section entitled **Weighting and Estimation** which contains four papers.

The first paper in this special section, by Singh and Mohl, gives an overview of calibration methods from a different perspective, with the objective of gaining a better heuristic understanding of these methods. Deville and Särndal presented calibration methods as minimizing the overall distance of the final weights from the survey weights, subject to the restriction that estimates of totals of certain covariates match known population totals. Singh and Mohl present different calibration methods as being derived from different models for the weight adjustment factors. Computational algorithms for different methods are provided in an appendix, and a numerical example is given to illustrate how the resulting weight adjustment factors might vary among the different methods.

Stukel, Hidioglou and Särndal also investigate calibration estimators, the class of design-based point estimators developed by Deville and Särndal. These estimators are derived from distance functions and allow for restricting of the final weights such that they are positive or upwardly bounded, thus avoiding the usual problem of negative weights that arises from using the regression estimator. Through simulation, the properties of a number of these estimators based on different distance functions are studied; particular emphasis is given to the properties of the corresponding variance estimators, specifically the Jackknife and the Taylor. The surprising conclusion is that the bias of both the point estimators and the corresponding variance estimators is minimal, even under severe restricting of the final weights.

Jayasuriya and Valliant compare three methods of deriving household weights for the Consumer Expenditure Survey of the U.S. Bureau of Labor Statistics. Survey weights are usually calibrated to population totals of individual level characteristics, resulting in different final weights for individuals in the same household. The principal person method defines the final weight for the household to be the same as that for a particular person in the household. The regression approach replaces the vector of auxiliary variables for each individual in a household by the household average, resulting in identical calibrated weights for persons in the same household. Another option is obtained by restricting the weight adjustment factors to avoid extreme or negative weights. Variations on these methods are compared with respect to the final weights and the estimated CVs for a variety of household expenditure categories.

In the final paper in the section on **Weighting and Estimation**, Chen and Chen consider the problem of confidence interval estimation for a finite population average when auxiliary information is available. Noting the earlier results of Royall and Cumberland that show that naive use of existing design-based methods results in confidence intervals with very poor conditional coverage probabilities, they suggest transformations of the data which improve the adherence to the underlying normality assumption and thus improve the coverage rates. Auxiliary information is incorporated in two ways: either directly into the inference when auxiliary information is known for each unit or through calibration with empirical likelihood when auxiliary information is known only at the population level. Through simulation applied to six real populations, they show that their methods perform well.

In their paper, Thompson and Fisher modify the one and two sample McNemar tests for use with complex survey data. They then apply the modified two sample test to data from the U.S. Bureau of the Census Current Population Survey's Split Panel Study to test whether or not the shift to computer assisted telephone interviewing using a redesigned questionnaire would affect the estimates of unemployment. Results of this test are discussed and compared to other research on the effect of CATI on unemployment estimates.

Eltinge and Jang suggest ways for evaluating the stability of estimates of variance components (specifically within-PSU variance estimators) and other related quantities, under a complex three-stage design. As measures, they consider a simple design-based variance estimator of the within-PSU variance estimator, as well as an estimated "degrees of freedom" approach. A simulation based method permits the assessment as to whether an observed stability measure is consistent with standard assumptions regarding variance estimator stability. They apply the proposed methods to NHANES III data and show that true stability properties may vary substantially across variables, and that within-PSU variance estimators can be substantially less stable than one would anticipate from using a simple count of secondary units within each stratum.

Berger discusses Chao's plan for sequentially selecting an unequal-probability sample of fixed size without replacement. In this context, he suggests an approximation of the second-order probabilities of inclusion in order to obtain an approximate estimator of the variance for the Horvitz and Thompson estimator. This variance is then compared to approximations given for other procedures or selection plans. Equivalence conditions for these approximations are presented.

Cowling, Chambers, Lindsay and Parameswaran present two techniques for producing spatially smoothed data and consider their implications in both small and large area estimation. For the small area application, the sample weights are spatially smoothed using a modified linear regression approach, which results in a decrease in the variance but an increase in the bias of the estimates. For the large area application, a nonparametric regression method is used to spatially smooth the data and then the smoothed data is mapped using a Geographic Information System package. The results of a simulation study are presented, in which the most appropriate method and level of smoothing for use in the maps is investigated.

Brick, Waksberg and Keeter suggest using information on interruptions of telephone service so as to adjust the survey estimates to compensate for undercoverage bias. The data collected on telephone service interruptions serve to reduce the bias, but at the same time the variance is likely to increase owing to the greater variability of the sampling weights. The results obtained from a national survey show a significant potential for reducing the mean square error of the estimates under certain conditions.

Finally, Pandher uses a model based approach to find an optimal partition of a survey population into take-all and take-some strata. The approach assumes that there is a single variable of interest and that probability proportional to size sampling is used in the take-some stratum. An algorithm is presented for determining the optimal cut point between the take-all and take-some groups. A key requirement for the algorithm is that the model expectation of the variance is a convex function of the number of units in the take-all stratum, which depends on the model assumptions and the form of the inclusion probabilities. The method is applied to Statistics Canada's Local Government Finance Survey.

The Editor



# Understanding Calibration Estimators in Survey Sampling

A.C. SINGH and C.A. MOHL<sup>1</sup>

## ABSTRACT

There exist well known methods due to Deville and Särndal (1992) which adjust sampling weights to meet benchmark constraints and range restrictions. The resulting estimators are known as calibration estimators. There also exists an earlier, but perhaps not as well known, method due to Huang and Fuller (1978). In addition, alternative methods were developed by Singh (1993), who showed that similar to the result of Deville-Särndal, all these methods are asymptotically equivalent to the regression method. The purpose of this paper is threefold: (i) to attempt to provide a simple heuristic justification of all calibration estimators (including both well known and not so well known) by taking a non-traditional approach; to do this, a model (instead of the distance function) for the weight adjustment factor is first chosen and then a suitable method of model fitting is shown to correspond to the distance minimization solution, (ii) to provide to practitioners computational algorithms as a quick reference, and (iii) to illustrate how various methods might compare in terms of distribution of weight adjustment factors, point estimates, estimated precision, and computational burden by giving numerical examples based on a real data set. Some interesting observations can be made by means of a descriptive analysis of numerical results which indicate that while all the calibration methods seem to behave similarly to the regression method for loose bounds, they however seem to behave differently for tight bounds.

**KEY WORDS:** Benchmark constraints; Distance minimization; Non-negative weights; Range restrictions.

## 1. INTRODUCTION

In providing estimates from sample surveys, sampling weights are commonly adjusted to obtain calibrated weights in order to match totals or benchmark constraints (BCs) for auxiliary variables. The methods of regression and raking are often used for this purpose. Although these methods have good asymptotic properties (see Deville and Särndal 1992), they may lead to calibrated weights with undesirable (finite sample) properties. The regression method can give negative weights while the raking procedure can produce very high weights. For this reason, range restrictions (RRs) may be imposed on the calibrated weights. It would be desirable to have a calibration method which (i) produces calibrated weights close to the original sampling weights; this can be achieved via minimization of a suitable distance function between the two sets of weights, (ii) meets BCs, and (iii) satisfies RRs. There exist several methods in the literature for weight adjustment under BCs and RRs, see *e.g.*, Deville and Särndal (1992, henceforth referred to as DS) for recent developments, and Huang and Fuller (1978) for earlier developments. For a review, as well as some further work, see Singh (1993, henceforth referred to as Singh). These methods are iterative in nature and can be classified into two families. Family I consists of methods which satisfy BCs after each iteration and continue to iterate until RRs are met. Family II, on the other hand, consists of methods which satisfy RRs after each iteration and continue to iterate until BCs are met.

Methods of DS belong to family II while that of Huang-Fuller belongs to family I. Two additional methods, one for each family, were proposed by Singh. Using arguments similar to DS, Singh extended the remarkable result of DS by showing that all of the methods in families I and II are asymptotically equivalent to the regression method.

In Section 2, a non-traditional approach is followed in introducing each method which is expected to help in understanding of calibration estimators. The functional form of the weight adjustment factor is first heuristically motivated and later on a connection between a suitable method of model fitting and minimization of the distance function is made. Alongside, computational algorithms are given as a quick reference for practitioners. A computer program in GAUSS software is available from the second author; see also Singh and Mohl (1997). In Section 3, numerical examples are presented to illustrate various methods using data from Statistics Canada's Family Expenditure (FAMEX) survey. It is of practical interest to see how different calibration methods might compare for a real data set. In particular, we examine by means of a descriptive analysis the impact of RRs on the computational burden, distribution of weight adjustment factors, point estimates and their variance. Related comparative studies on calibration methods based on real data sets are due to Deville, Särndal and Sautory (1993) and Stukel and Boyer (1993). These studies, however, are restricted to family II methods and are primarily concerned with the distribution of weight adjustment factors. Finally, Section 4 contains a discussion.

<sup>1</sup> A.C. Singh, Methodology Research Advisory Group, and C.A. Mohl, Health Statistics Methods Section, Household Survey Methods Division, Statistics Canada, Ottawa, K1A 0T6.

## 2. HEURISTIC JUSTIFICATION OF CALIBRATION ESTIMATORS

We will use the following notation. Let  $n, N$  denote respectively the sample size and the population size. Let  $h_k$  denote the initial or  $h$ -weight (used in the expansion or Horvitz-Thompson estimator  $\sum_{k=1}^n y_k h_k$ ) for the  $k$ -th element where  $y_k$  is the value of the study variable. It is assumed that the  $h$ -weights include adjustments for any non-response. The parameter of interest is the population total for  $y$ , denoted by  $\tau_y$ . For each  $k$ , there are  $p$ -auxiliary variables,  $x_{kj}, j = 1, \dots, p$  for which the population total or benchmark constraint,  $\tau_{xj} = \sum_{k=1}^N x_{kj}$  for each  $j$  is assumed to be known. The transposed  $p$ -vector  $\mathbf{x}_k$  denotes  $(x_{k1}, \dots, x_{kp})$ , the  $k$ -th row of the  $n \times p$  matrix  $\mathbf{X}$ . Let  $c_k^{(v)}$  denote the calibrated or  $c$ -weight for the  $k$ -th element at the  $v$ -th iteration. At  $v = 0$ ,  $c_k^{(v)} = h_k$ . The expansion estimators of population totals for variables  $y$  and  $x_j$  using  $c$ -weights at the  $v$ -th iteration are denoted by  $\hat{\tau}_y^{(v)}$  and  $\hat{\tau}_{xj}^{(v)}$  respectively.

The RRs are specified by the condition  $L \leq g_k \leq U$  where  $g_k = c_k/h_k$  and  $L < 1 < U$ , where  $L$  and  $U$  denote suitable lower and upper bounds. The adjustment factors (*i.e.*,  $g_k$ 's) are also called  $g$ -weights. First we consider the unrestricted case (*i.e.*, calibration without RRs) and then the restricted case. All methods in the restricted case require iterations for finding a solution. It is assumed that the iterative process converges in a finite number of iterations.

The criterion for convergence is defined as follows. For the iterative process to meet RRs, a tolerance level  $\epsilon$  (*e.g.*, .005 or .01) for family I is defined so that the process terminates if the maximum absolute relative error (ARE) for RRs is  $\leq \epsilon$ . Similarly, a tolerance level ( $\delta > 0$ ) for family II is defined for meeting BCs by iterations. The reason for this is that our primary goal is not minimization of the distance function, but to find a solution which satisfies BCs and RRs. In addition to  $\epsilon$  and  $\delta$ , a parameter  $v_{\max}$  is defined which limits the number of iterations.

There are seven methods considered in this paper, two for the unrestricted case, two for restricted case in family I and the remaining three also for the restricted case but in family II. We have given alternative names to existing methods to facilitate understanding of the relationship between different methods. The naming convention is based on the well known distance measures used in the analysis of count data.

Note that since all the methods are asymptotically equivalent to the regression method, the asymptotic variance of  $\hat{\tau}_y$  can be estimated for each method by  $\sum_k \sum_l (\pi_{kl} - \pi_k \pi_l) \pi_{kl}^{-1} (e_k g_k)(e_l g_l)$ , as in DS (equation 3.4) where  $\pi_k, \pi_{kl}$  are respectively the first and second order inclusion probabilities,  $e_k$  are the sample residuals  $y_k - \hat{\mathbf{B}}' \mathbf{x}_k$  with  $\hat{\mathbf{B}}' = (\mathbf{y}' \Gamma_0 \mathbf{X}) (\mathbf{X}' \Gamma_0 \mathbf{X})^{-1}$ , and  $\Gamma_0$  is the  $n \times n$  matrix  $\text{diag}(\mathbf{h})$ .

### 2.1 METHOD 1 (Linear Regression or Unrestricted Modified Chi Square, MCS-u)

This method is the simplest and gives rise to the popular generalized regression estimator of Särndal (1980). Here, the

model for the adjustment factor is taken to be linear in  $\mathbf{x}$ , *i.e.*,  $g_k = 1 + \mathbf{x}_k' \boldsymbol{\lambda}$ , for some  $p$ -vector of model parameters  $\boldsymbol{\lambda}$  which satisfies BCs. That is,  $\sum_{k=1}^n h_k (1 + \mathbf{x}_k' \boldsymbol{\lambda}) x_{kj} = \tau_{xj}$ , for all  $j$ . This gives rise to  $\boldsymbol{\lambda}^{\text{MCS-u}}$  as  $(\mathbf{X}' \Gamma_0 \mathbf{X})^{-1} (\tau_{\mathbf{x}} - \hat{\tau}_{\mathbf{x}}^{(0)})$ . The  $c$ -weights remain close to the  $h$ -weights in the sense that the above choice of  $g$ -weights minimizes the distance function,  $\Delta^{\text{MCS-u}}(\mathbf{c}, \mathbf{h}) = \sum_{k=1}^n (c_k - h_k)^2 / h_k$  subject to BCs. Note that the  $g$ -weights could be negative for some  $k$ . This is rather undesirable in practice although the simplicity of the method is quite attractive. The computational algorithm for MCS-u is given in Appendix A1.

### 2.2 METHOD 2 (Raking or Unrestricted Modified Discrimination Information, MDI-u)

This method is also commonly used. Here, the model for the adjustment factor  $g_k$  is taken as  $\exp(\mathbf{x}_k' \boldsymbol{\lambda})$ , thus making it necessarily non-negative. Unlike the case of method 1, the model parameter vector  $\boldsymbol{\lambda}^{\text{MDI-u}}$  is obtained iteratively to meet BCs. The iterations can be started with  $\boldsymbol{\lambda}^{\text{MCS-u}}$  from the GR-estimator, *i.e.*, for iteration 1, set  $\boldsymbol{\lambda}^{(1)} = \boldsymbol{\lambda}^{\text{MCS-u}}$ , which implies  $c_k^{(1)} = h_k \exp(\mathbf{x}_k' \boldsymbol{\lambda}^{(1)})$ . These  $c$ -weights, in general, do not satisfy BCs. For iteration 2 of this method, the  $\boldsymbol{\lambda}^{(1)}$  is adjusted (by a term of smaller order) to define  $\boldsymbol{\lambda}^{(2)}$  as  $\boldsymbol{\lambda}^{(1)} + (\mathbf{X}' \Gamma_1 \mathbf{X})^{-1} (\tau_{\mathbf{x}} - \hat{\tau}_{\mathbf{x}}^{(1)})$ , where  $\Gamma_1 = \text{diag}(\mathbf{c}^{(1)})$ . The  $\boldsymbol{\lambda}$  term is defined similarly for further iterations until convergence, *i.e.*, until BCs are met. The  $c$ -weights remain close to  $h$ -weights because iterations used in the above method constitute the Newton-Raphson steps for minimizing the distance function,  $\Delta^{\text{MDI-u}}(\mathbf{c}, \mathbf{h}) = \sum_{k=1}^n [c_k \log(c_k/h_k) - c_k + h_k]$  subject to BCs. Note that although the  $g$ -weights are non-negative, they could be very high which is clearly not desirable in practice. The computational algorithm for MDI-u is given in Appendix A2.

### 2.3 METHOD 3 (Modified Huang-Fuller or Scaled Modified Chi Square, SMCS)

This method belongs to family I of the restricted case and is a slight modification of the method due to Huang and Fuller as given in Singh; see also Fuller, Loughin, and Baker (1994). As in regression, the model for the adjustment factor is taken to be linear in  $\mathbf{x}$ . To facilitate the satisfaction of RRs by these adjustments, a scaling factor  $q_k$  ( $0 < q_k \leq 1$ ), is used for each  $k$  so that the change in  $h$ -weights for those units whose  $g_k$ 's tend to go outside the bounds  $[L, U]$  is reduced. Thus, the  $g$ -weight is given by  $g_k = 1 + q_k \mathbf{x}_k' \boldsymbol{\lambda}$  where the model parameters  $\mathbf{q}$  and  $\boldsymbol{\lambda}$  are chosen iteratively in the sense that  $\boldsymbol{\lambda}$  is found for a given  $\mathbf{q}$  and then  $\mathbf{q}$  is found for a given  $\boldsymbol{\lambda}$ . We start with  $q_k^{(0)} = 1$  for all  $k$  and set  $\boldsymbol{\lambda}^{(1)} = \boldsymbol{\lambda}^{\text{MCS-u}}$  for iteration 1. Now, clearly  $\mathbf{c}^{(1)}$  satisfies BCs but RRs need not be satisfied. Depending on the location of  $g_k$ 's in relation to  $[L, U]$ , a working rule can be used to define  $q_k$ 's so that the  $q_k$ 's discount more for those units which are farther outside of the boundaries than those which are nearer. The scaling factors  $q_k^{(1)}$  so determined, define in turn  $\boldsymbol{\lambda}^{(2)}$  for iteration 2 as  $(\mathbf{X}' \Gamma_1 \mathbf{X})^{-1} (\tau_{\mathbf{x}} - \hat{\tau}_{\mathbf{x}}^{(1)})$  where  $\Gamma_1 = \text{diag}(q_k^{(1)} h_k)$ ,  $q_k^{[1]} = q_k^{(0)} q_k^{(1)}$ ,



$\lambda^{(2)}$  satisfying BCs after the iteration. Note that under usual regularity conditions,  $\lambda^{(2)}$  differs from  $\lambda^{(1)}$  only by a term of smaller order, since the maximum absolute difference  $|q_k^{(1)} - 1|$  is small. Next, if  $c^{(2)}$  after iteration 2 does not satisfy RRs, the scaling factors  $q_k^{(2)}$  are defined appropriately and compounded with  $q_k^{(1)}$  to get  $q_k^{(2)}$  for use in iteration 3. The  $\lambda^{(3)}$  for iteration 3 is then obtained as before so that BCs are satisfied after the iteration. Iterations continue until convergence, *i.e.*, until RRs are met. The weight vector  $c^{\text{SMCS}}$  is close to  $h$  because at each iteration  $v$ ,  $c^{(v)}$  minimizes the distance function  $\Delta_v^{\text{SMCS}}(c, h) = \sum_{k=1}^n (c_k - h_k)^2 / h_k q_k^{[v-1]}$  subject to BCs, where  $q_k^{[v-1]} = q_k^{(0)} q_k^{(1)} \dots q_k^{(v-1)}$  for  $v \geq 1$ . Note that unlike the previous methods, the distance function varies from iteration to iteration.

The computational algorithm for SMCS is given in Appendix A3. Note that in the algorithm,  $[L, U]$  is shrunk to  $[L', U']$  by means of a parameter  $\alpha$  where  $L' = \alpha L + 1 - \alpha$ ,  $U' = \alpha U + 1 - \alpha$ , and  $0 < \alpha \leq 1$ . This implies that some units that are inside  $[L, U]$  but close to the boundary are also discounted. This helps to speed up the convergence. Another parameter  $\beta$ ,  $0 \leq \beta \leq 1$  is also introduced to allow differential discounting of different units.

## 2.4 METHOD 4 (Shrinkage-Minimization, SM)

This method also belongs to family I and is due to Singh. As in regression, the model for the adjustment factor is taken to be linear in  $x$ , but a new parameter termed the shrinkage factor  $\psi_k$  ( $0 < \psi_k \leq 1$ ) is used for each  $k$  so that  $g_k$ 's meet RRs, *i.e.*,  $g_k$  is set at  $(1 + \psi_k x_k' \lambda(k))$ . Notice that  $\lambda$  is allowed to depend on  $k$  through  $\psi_k$  and  $x_k$ . Unlike SMCS, here the  $g$ -weights, after discounting, satisfy RRs exactly, *i.e.*, those  $g$ -weights which are outside  $[L, U]$  are shrunk to lie on or inside the boundary. Therefore,  $\psi_k$ 's can be defined quite easily in practice. The model parameters  $\psi$  and  $\lambda$  are chosen iteratively in a manner analogous to that for SMCS. We start with  $\psi_k^{(0)} = 1$  and set  $\lambda^{(1)} = \lambda^{\text{MCS-u}}$  for iteration 1 to obtain  $g_k^{(1)}$  as  $(1 + \psi_k^{(0)} x_k' \lambda^{(1)})$ . Clearly BCs are satisfied after the iteration but RRs need not be. Before iteration 2,  $g_k^{(1)}$  is shrunk by  $\psi_k^{(1)}$  to obtain  $g_k^{(1)*}$  as  $(1 + \psi_k^{(1)} x_k' \lambda^{(1)})$  where  $\psi_k^{(1)} = \psi_k^{(0)} \phi_k^{(1)}$ , which meets RRs. Given  $\psi^{(1)}$ ,  $\lambda^{(2)}(k)$  is obtained as  $\lambda^{(1)} + (1/\psi_k^{(1)})(X' \Gamma_1 X)^{-1} (\tau_x - \hat{\tau}_x^{(1)*}) + x_k'(X' \Gamma_1 X)^{-1} (\tau_x - \hat{\tau}_x^{(1)*}) \lambda^{(1)}$  where  $\Gamma_1 = \text{diag}(c^{(1)*})$ ,  $c_k^{(1)*} = h_k g_k^{(1)*}$ , and  $\hat{\tau}_x^{(1)*}$  is the expansion estimator using  $c^{(1)*}$ -weights. Again BCs are satisfied after the iteration but RRs need not be. Note that  $\lambda^{(2)}(k)$  differs from  $\lambda^{(1)}$  by a term of smaller order uniformly over  $k$ . Iterations are continued until convergence, *i.e.*, until RRs are met. The weight vector  $c^{\text{SM}}$  is close to  $h$  because at each iteration  $v \geq 1$ ,  $c^{(v)}$  minimizes the distance function,  $\Delta_v^{\text{SM}}(c, c^{(v-1)*}) = \sum_{k=1}^n (c_k - c_k^{(v-1)*})^2 / c_k^{(v-1)*}$  subject to BCs. Note that in practice  $c^{(v)*}$  can be obtained directly from  $c^{(v)}$  without having to calculate  $\psi^{(v)}$  separately. As with SMCS, the distance function depends on the iteration.

The computational algorithm is given in Appendix A4. Recall that in the above method, if a  $g$ -weight falls outside of the  $L$  and  $U$  boundaries, an adjustment is made to bring the  $g$ -weight back to the  $L$  or  $U$  boundary. A new parameter

$\alpha$  ( $0 < \alpha \leq 1$ ) is introduced to allow the user to bring the  $g$ -weight farther inside the boundary to a point  $L'$  or  $U'$  ( $L' = \alpha L + 1 - \alpha$ ,  $U' = \alpha U + 1 - \alpha$ ). This is somewhat similar to the  $\alpha$  parameter of SMCS. Another parameter  $\eta$  ( $0 < \eta \leq \alpha \leq 1$ ) is introduced to adjust the  $g$ -weights to the level  $L'$  or  $U'$  also for those units which are within  $[L, U]$ , but close to the boundary in that they are outside  $[L'', U'']$  where  $L'' = \eta L + 1 - \eta$ ,  $U'' = \eta U + 1 - \eta$ . All these parameters help speed up the convergence in general.

## 2.5 METHOD 5 (Linear Truncated or Restricted Modified Chi Square, MCS-r)

This well known method belongs to family II of the restricted case and is due to DS. As in SM, the model for the adjustment factor is taken to be linear in  $x$  with a new parameter termed the truncation factor  $\phi_k$  ( $0 < \phi_k \leq 1$ ) which is used for each  $k$  so that  $g_k$ 's meet RRs, *i.e.*,  $g_k$  is set at  $(1 + \phi_k x_k' \lambda(k))$ . The only difference between the truncation factor  $\phi_k$  used here and the shrinkage factor used in SM is that here those  $g$ -weights which are outside  $[L, U]$  are always adjusted to lie exactly on the boundary. The model parameters are chosen iteratively. Initially we set  $\phi_k^{(0)} = 1$  and at iteration 1,  $\lambda^{(1)} = \lambda^{\text{MCS-u}}$  to obtain  $\tilde{g}_k^{(1)} = (1 + \phi_k^{(0)} x_k' \lambda^{(1)})$ , which is further adjusted (or truncated) to obtain  $g_k^{(1)}$  as  $(1 + \phi_k^{(1)} x_k' \lambda^{(1)})$  where  $\phi_k^{(1)} = \phi_k^{(0)} \phi_k^{(1)}$ , so that RRs are met. However,  $g^{(1)}$  may not satisfy BCs. Note that the difference between  $g^{(1)}$  and  $g^{\text{MCS-u}}$  is of smaller order. Now, for iteration 2,  $\lambda^{(1)}$  is adjusted by a term of smaller order (uniformly over  $k$ ) to define  $\lambda^{(2)}(k)$  as  $\lambda^{(1)} + (1/\phi_k^{(1)})(X' \Gamma_1 X)^{-1} (\tau_x - \tau_x^{(1)})$ , where  $\Gamma_1 = \text{diag}(h)$  except that the diagonal elements are truncated to zero for all those  $k$  for which  $\phi_k^{(1)} < 1$ , *i.e.*, those units which were truncated at the previous iteration. This discounting of diagonal elements is somewhat similar to using a zero scaling factor in SMCS. In the second iteration, we have  $\tilde{g}_k^{(2)} = 1 + \phi_k^{(1)} x_k' \lambda^{(2)}(k)$  and the truncation factors  $\phi_k^{(2)}$  are used to obtain  $g_k^{(2)}$  which satisfy RRs. The successive iterations are defined in a similar manner. Clearly, unlike SM, here RRs are met at each iteration. Iterations are continued until BCs are met. The weight vector,  $c^{\text{MCS-r}}$  is close to  $h$  because the iterations defined above constitute the Newton-Raphson steps for minimizing the distance function  $\Delta^{\text{MCS-r}}(c, h) = \sum_k (c_k - h_k)^2 / h_k$  if  $Lh_k \leq c_k \leq Uh_k$ ;  $\infty$  otherwise, subject to BCs. The computational algorithm is given in Appendix A5. Note that, in practice, it is more convenient to work with  $g_k^{(v)}$  directly without having to compute  $\phi_k^{(v)}$  separately.

## 2.6 METHOD 6 (Restricted Modified Discrimination Information or MDI-r)

This method also belongs to family II and was proposed by Singh following the lines of DS in developing MCS-r. It is related to MDI-u in the same way as MCS-r is to MCS-u. The basic idea is to adjust parameters  $\phi$  and  $\lambda$  in the adjustment factor  $g_k = \phi_k \exp(x_k' \lambda)$  so that RRs and BCs are satisfied. The truncation parameter  $\phi$  is similar to that for MCS-r. This



is done iteratively. Similar to MCS-r, at iteration 1 we set  $\hat{g}_k^{(1)} = \phi_k^{(0)} \exp(\mathbf{x}_k' \lambda^{(1)})$  where  $\phi_k^{(0)} = 1, \lambda^{(1)} = \lambda^{\text{MCS-u}}$ , which is further adjusted by a term of smaller order to obtain  $\hat{g}_k^{(1)}$  as  $\phi_k^{(1)} \exp(\mathbf{x}_k' \lambda^{(1)})$  so that RRs are met, i.e., it lies in  $[L, U]$ . Next for iteration 2,  $\hat{g}_k^{(1)}$  is adjusted by a term of smaller order to obtain  $\hat{g}_k^{(2)}$  as  $\phi_k^{(2)} \exp(\mathbf{x}_k' \lambda^{(2)})$ , where  $\lambda^{(2)} = \lambda^{(1)} + (\mathbf{X}' \Gamma_1 \mathbf{X})^{-1} (\tau_x - \hat{\tau}_x^{(1)})$ , and  $\Gamma_1 = \text{diag}(h_k \hat{g}_k^{(1)})$  except that the diagonal elements are truncated to 0 for all those  $k$  for which  $\phi_k^{(1)} < 1$ . The truncation factors  $\phi_k^{(2)}$  are used to ensure that RRs are met. Iterations are continued until convergence, i.e., until BCs are met. The weight vector  $\mathbf{c}^{\text{MDI-r}}$  is close to  $\mathbf{h}$  because the iterations defined above constitute the Newton-Raphson steps for minimizing the distance function  $\Delta^{\text{MDI-r}}(\mathbf{c}, \mathbf{h}) = \sum_{k=1}^n [c_k \log(c_k/h_k) - c_k + h_k]$  if  $Lh_k \leq c_k \leq Uh_k$ ;  $\infty$  otherwise, subject to BCs. Note that in practice, the truncation factors are not needed separately to compute  $\hat{g}_k^{(v)}$ . Appendix A6 gives the computational algorithm for MDI-r.

## 2.7 METHOD 7 (Logit or Generalized Modified Discrimination Information, GMDI)

This is the last method considered. This well known method of family II is due to DS. As in the raking method, we start with  $\exp(\mathbf{x}_k' \lambda)$  and an inverse logit-type transformation is used to ensure that the adjustment factor satisfies RRs. The model for the adjustment factor is given by  $g_k = [(U - 1) + (1 - L) \exp(A\mathbf{x}_k' \lambda)]^{-1} [L(U - 1) + U(1 - L) \exp(A\mathbf{x}_k' \lambda)]$ , where  $A = (1 - L)^{-1} (U - 1)^{-1} (U - L)$ . This adjustment factor, unlike other methods, lies necessarily inside the interval  $[L, U]$ , i.e., does not take boundary values. As  $L \rightarrow 0$  and  $U \rightarrow \infty$ , the factor reduces to the familiar inverse logit form,  $\exp(\mathbf{x}_k' \lambda) / [1 + \exp(\mathbf{x}_k' \lambda)]$ . The model parameter  $\lambda$  is obtained iteratively to meet BCs. Starting with  $\lambda^{\text{MCS-u}}$  as  $\lambda^{(1)}$  for iteration 1, we adjust by a smaller order term to obtain  $\lambda^{(2)}$  as  $\lambda^{(1)} + (\mathbf{X}' \Gamma_1 \mathbf{X})^{-1} (\tau_x - \hat{\tau}_x^{(1)})$  where  $\Gamma_1 = \text{diag}(h_k d_k^{(1)})$ ,  $d_k^{(1)} = (U - 1)^{-1} (1 - L)^{-1} (U - g_k^{(1)}) (g_k^{(1)} - L)$ . Further iterations are done in a similar manner until BCs are met. The weight-vector  $\mathbf{c}^{\text{GMDI}}$  is close to  $\mathbf{h}$  in the sense that subject to BCs, the above iterative process corresponds to the Newton-Raphson algorithm for minimizing the distance function  $\Delta^{\text{GMDI}}(\mathbf{c}, \mathbf{h})$  given by  $A^{-1} \sum_{k=1}^n h_k [(g_k - L) \log\{(1 - L)^{-1} (g_k - L)\} + (U - g_k) \log\{(U - 1)^{-1} (U - g_k)\}]$ . Appendix A7 gives the computational algorithm for GMDI.

## 3. NUMERICAL EXAMPLES

### 3.1 Data Description

We consider application of the seven adjustment methods described above to data from the 1990 Statistics Canada's Family Expenditure (FAMEX) Survey for the two cities (or domains) of Regina and Saskatoon in the province of Saskatchewan. Four study variables are considered: annual expenditures on owned dwelling for repair and renovation, furniture and equipment, ladies' clothing, and men's clothing. The FAMEX survey is a supplementary survey to the Canadian Labour Force Survey (LFS) and, therefore, is based on the LFS design – a multistage stratified cluster sample of

households, see Singh *et al.* (1990). Samples are drawn independently from the two cities of Regina and Saskatoon. Respectively for the two cities, the numbers of strata are 30 and 34, and the numbers of primary sampling units (PSUs) selected in the sample are 111 and 94. The total numbers of sampled households are 321 and 278, while the corresponding numbers ( $n$ ) of individuals are 797 and 712.

### 3.2 Benchmark Constraints, Range Restrictions and Common Weights per Household

The number ( $p$ ) of BCs is four for each domain. They correspond to the demographic population counts for the four groups: age  $< 15$ , age  $\geq 15$ , one person households, and households with two or more persons. The corresponding counts are 40696, 139047, 12746, and 48457 for Regina, and 42544, 139299, 20628, and 52059 for Saskatoon. Thus, the total numbers of households for the two domains are 61203 and 72687 respectively and the corresponding population sizes ( $N$ ) are 179743 and 181843. The auxiliary  $x$ -variables here are indicators for the above four groups.

For Regina, (min, max) of  $g$ -weights are obtained as  $(-0.72, 2.74)$  and  $(0.19, 3.95)$  respectively for regression and raking methods. It is therefore of interest to make them nonnegative for regression and to reduce the high weights for raking. Two types of RRs are chosen: one has somewhat loose bounds with  $L = 1/5$  and  $U = 5$  and the other has somewhat tight bounds with  $L = 2/5$  and  $U = 5/2$ . For Saskatoon, (min, max) of  $g$ -weights are obtained as  $(0.86, 1.08)$  and  $(0.87, 1.09)$  respectively for regression and raking methods. Note that both methods give  $g$ -weights close to 1, and therefore there is no real need for RRs. However, for the sake of illustration, we choose  $L = 0.88$  and  $U = 1.12$ .

The initial sampling weights or  $h$ -weights of individuals in the same household are common and equal to the weight of that household. It is desirable that after calibration, all members of a household have the same  $c$ -weights. This can be achieved by modifying the  $\mathbf{X}$  matrix so that  $x_j$ -values for each person in the same household are common and equal to the average value for the household, see, e.g., Lemaître and Dufour (1987). We also perform an initial scaling on the  $h$ -weights so that they add up to  $N$ ; this is similar to the Hájek modification of the Horvitz-Thompson estimator. This scaling essentially redefines  $[L, U]$  to make them meaningful for calibration of  $h$ -weights.

### 3.3 Descriptive Measures for Comparison

For comparing various methods, we consider four types of descriptive measures:

- Summary statistics for the distribution of the  $g$ -weights,
- Point estimates for several variables,
- Estimated precision of the calibration estimates, and
- Computational burden imposed by each method.

The first measure consists of a graphical summary using a box plot for  $g$ -weights, and the standard deviation of  $g$ -weights,  $\text{SD}(g)$ , defined as  $[N^{-1} \sum_{k=1}^n h_k (g_k - 1)^2]^{1/2}$ . Note



that the mean of  $g$ -weights, *i.e.*,  $N^{-1} \sum_{k=1}^n h_k g_k$ , is 1 in view of the fact that  $\sum h_k = \sum c_k = N$ , and the  $SD(g)$  also equals  $[N^{-1} \sum_{k=1}^n (c_k - h_k)^2 / h_k]^{1/2}$ , the square root of a normalized chi-square type distance for measuring closeness between  $h$ - and  $c$ -weights. For comparing point estimates and their precision for estimating parameter for each variable  $y$  of interest, we compute relative difference (RD) and relative precision (RP) with respect to the MCS- $u$  weights, *i.e.*, relative to the regression estimator. Denoting an estimator based on  $c$ -weights as a  $c$ -estimator, we have RD as ( $c$ -estimator minus regression estimator) divided by the regression estimator, and RP as  $SE(\text{regression estimator})$  divided by  $SE(c\text{-estimator})$ . Note that for the numerical examples under consideration, variances are computed using jackknifing by deleting PSUs. Finally, the computational burden is expressed in terms of the number of iterations. Testing has shown that for all the restricted methods, each iteration takes a similar amount of time and hence a good comparison of their computational burden is the number of iterations required for convergence.

### 3.4 Specification of Other Parameters

We also need to specify some other parameters, namely,  $\alpha$ ,  $\beta$  for SMCS, and  $\alpha$ ,  $\eta$  for SM. Empirically, values of  $\alpha = 0.67$ ,  $\eta = 0.9$  and  $\beta = 0.8$  are found to perform well. The tolerance levels  $\epsilon$  for family I and  $\delta$  for family II are set at 0.01, and  $v_{\max}$  is set at 10.

## 3.5 Results: A Descriptive Analysis

### 3.5.1 Distribution of $g$ -weights

We first consider the Regina data. Figure 1 gives a box plot of the distribution of  $g$ -weights with  $L = 0.4$  and  $U = 2.5$ . Note that there are negative  $g$ -weights (and hence negative  $c$ -weights) for MCS- $u$  and large  $g$ -weights (which produce large  $c$ -weights) for the MDI- $u$  method. For MCS- $u$ , the fraction of  $g$ -weights  $< 0$  is 4.9%, the fraction  $< 0.4$  is 5.9%, the fraction above 2.5 is 1.25% while above 3.5 is 0%. For MDI- $u$ , the fraction below 0.4 is 4.9%, the fraction  $> 2.5$  is 4.3% and above 3.5 is 1.25%. Thus, both methods yield  $c$ -weights which are out of bounds with respect to RRs with tight bounds. The range restricted methods all have median  $g$ -weights between 0.65 and 0.75; the SMCS  $g$ -weights show, however, the most clustering around the median. Table 1 shows that under loose bounds, the  $SD(g)$  for each restricted method is slightly higher (about 7%) than the regression method, but for tight bounds, the difference increases to about 15% for family I and about 10% for family II.

Now for the Saskatoon data, Figure 2 gives a box plot of  $g$ -weights with  $L = 0.88$  and  $U = 1.12$ . For both regression and raking methods, about 5.6% are below  $L$  and 0% are above  $U$ . All methods have similar interquartile range for  $g$ -weights with medians slightly above 1. Also it is seen from Table 1 that  $SD(g)$  for all the methods (restricted and unrestricted) are about the same and quite small.

**Table 1**  
Number of Iterations and  $SD(g)$   
( $\alpha = .67$ ,  $\beta = .8$ ,  $\eta = .9$ ,  $\epsilon = \delta = .01$ ,  $v_{\max} = 10$ )

Method	Regina				Saskatoon	
	$L = 0.2, U = 5.0$ (Loose bounds)		$L = 0.4, U = 2.5$ (Tight bounds)		$L = 0.88, U = 1.12$	
	Number of iterations	$SD(g)$	Number of iterations	$SD(g)$	Number of iterations	$SD(g)$
Family I						
SMCS	2	0.647	3	0.702	2	0.071
SM	2	0.636	4	0.689	2	0.070
Family II						
MCS-r	2	0.628	3	0.654	1	0.069
MDI-r	3	0.642	3	0.660	1	0.069
GMDI	3	0.640	3	0.659	2	0.069

**Note:** For the unrestricted (or no bounds) case, the number of iterations and  $SD(g)$  are: for Regina MCS- $u$  and MDI- $u$  are (1,0.599) and (3,0.647) respectively; for Saskatoon MCS- $u$  and MDI- $u$  are (1,0.070) and (1,0.069) respectively.

### 3.5.2 Relative Difference of Point Estimates

Tables 2(a) and (b) show that for Regina, under loose bounds RD is small for all the methods for each of the variables. In fact, it is negligible except for the variable "owned dwelling" for which it is generally under 4%. However, under tight bounds, it increases somewhat but remains small with values ranging between 1% and 5%. For Saskatoon (Table 2c), under the given bounds RD is negligible for all the methods.

### 3.5.3 Estimated Relative Precision of Estimates

For Regina, under loose bounds, RP is generally within 5% (of the precision of the regression estimator) for all methods and all variables except for MDI-r with the variable "ladies' clothing" for which it is lower by 9%. However, under tight bounds, RP varies more and is now generally within 9% except for SMCS and SM with the variable "Men's clothing" (RP is lower by 20%) and MDI-r for the variable "Ladies' clothing" for which RP is lower by 11%. For Saskatoon (Table 2c), under the chosen bounds RP is close to 1 for all cases.

### 3.5.4 Computational Burden

For Regina (Table 1), under loose bounds each method takes two or three iterations. As the bounds are tightened, most of the methods require more iterations to converge. To see how tightly the bounds could be squeezed before encountering convergence problems, three more sets of bounds were used with  $[L, U] = [0.425, 2.35]$ ,  $[0.45, 2.22]$  and  $[0.475, 2.11]$ . These results are not shown in the table. With  $v_{\max}$  as 10, the SM method does not converge for  $[0.425, 2.35]$ . The SMCS and GMDI methods do not converge for  $[0.45, 2.22]$  and the MCS-r and MDI-r finally have

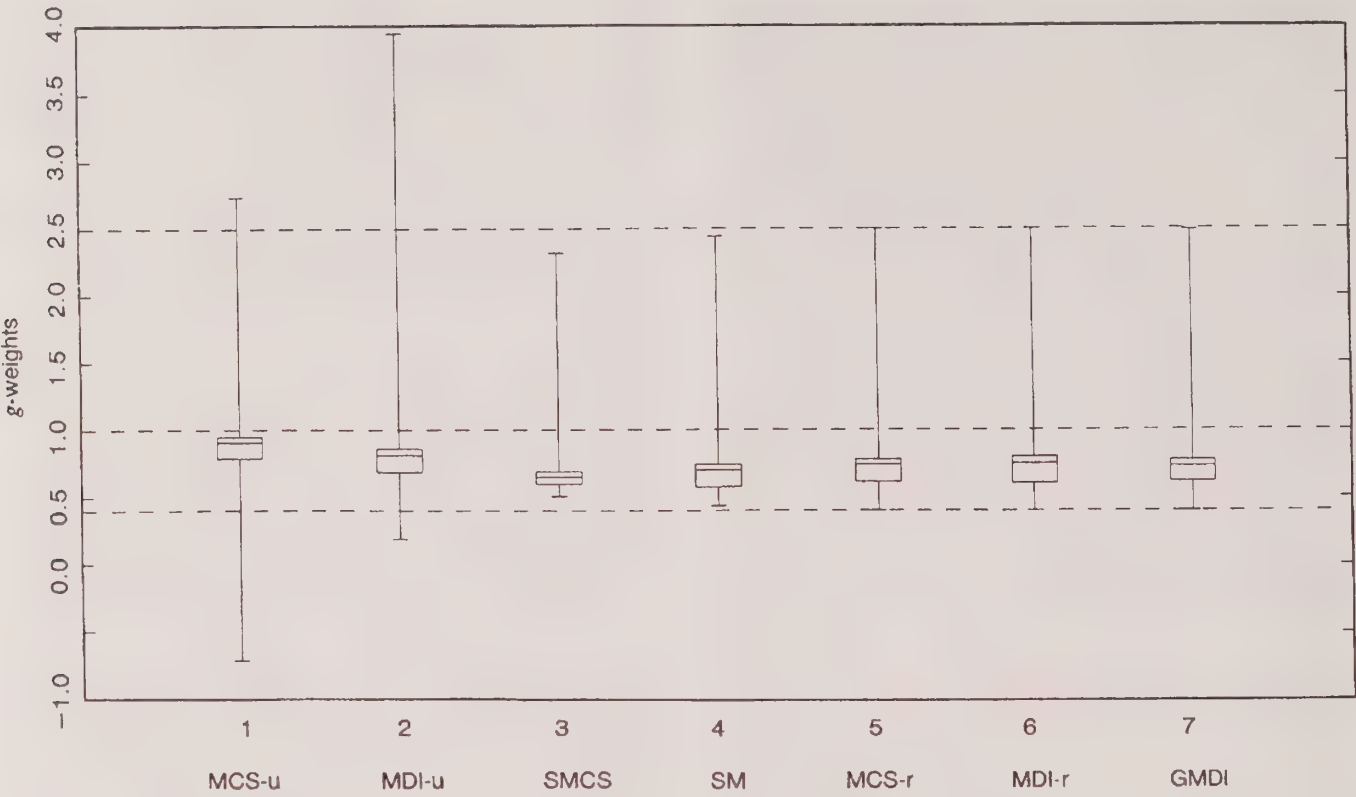


Figure 1. Box Plot: g-weights for Regina FAMEx data ( $L = 0.4, U = 2.5$ )

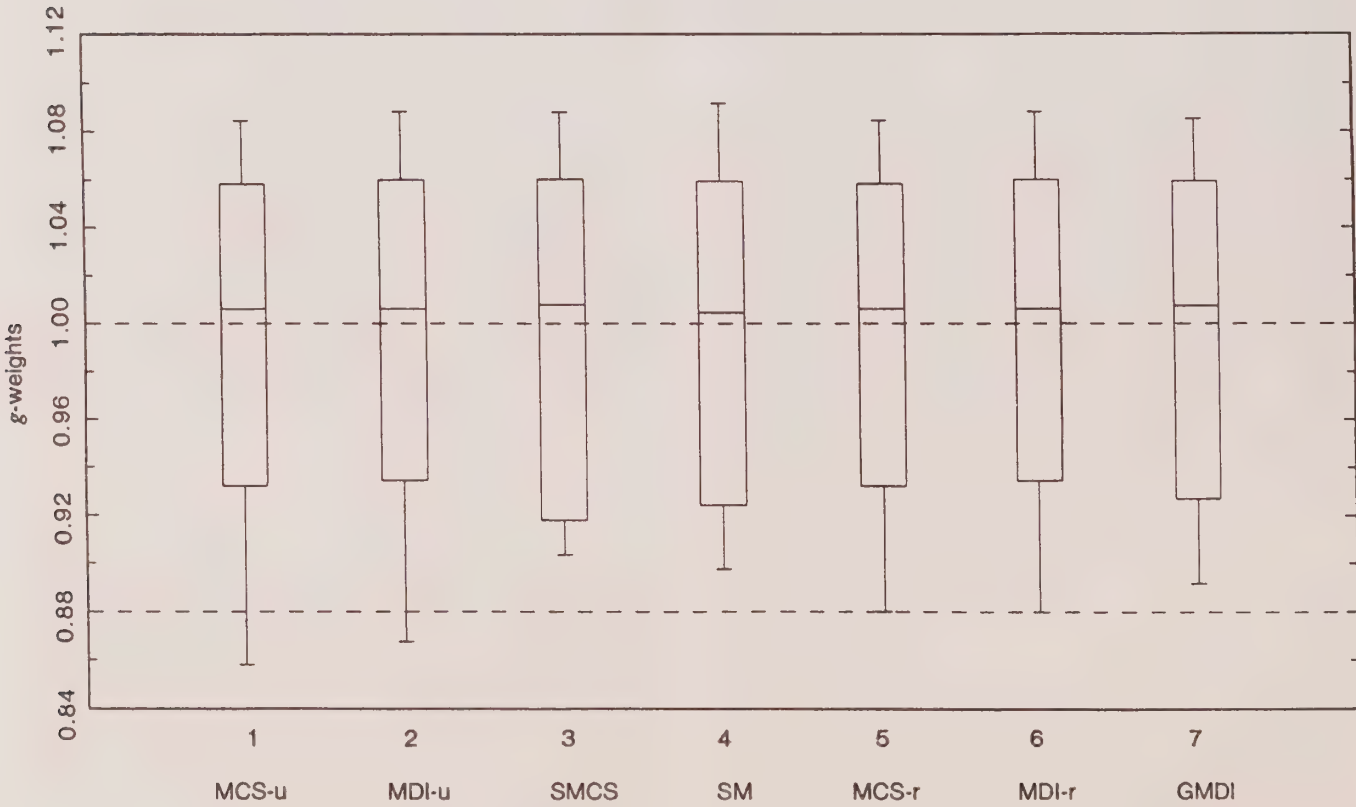


Figure 2. Box Plot: g-weights for Saskatoon FAMEx data ( $L = 0.88, U = 1.12$ )



Table 2a

Difference in Point Estimates and Precision Relative to Regression Estimator ( $\alpha = .67, \beta = .8, \eta = .9, \epsilon = \delta = .01, v_{\max} = 10$ )  
Regina:  $L = 0.2, U = 5.0$  (Loose Bounds)

	Owned Dwelling		Furniture\Equipment	
	RD	RP	RD	RP
Family I				
SMCS	-0.043	1.047	0.001	1.032
SM	-0.036	1.032	-0.002	1.040
Family II				
MCS-r	-0.032	1.035	0.002	1.034
MDI-r	-0.033	0.991	-0.008	1.037
GMDI	-0.037	0.999	-0.004	1.041
	Ladies' Clothing		Men's Clothing	
Family I				
SMCS	0.015	0.931	0.009	0.952
SM	0.010	0.951	0.006	0.968
Family II				
MCS-r	0.011	0.950	0.008	0.964
MDI-r	0.007	0.911	-0.001	0.961
GMDI	0.009	0.940	0.002	0.968

## Notes:

1. RD and RP denote respectively "relative difference" and "relative precision".
2. For the unrestricted (or no bounds) case, the corresponding measures for the raking (MDI-u) method relative to regression are (-0.034, 1.005), (-0.008, 1.049), (0.004, 0.968) and (0.002, 0.980) for the four study variables respectively.

Table 2b

Difference in Point Estimates and Precision Relative to Regression Estimator ( $\alpha = .67, \beta = .8, \eta = .9, \epsilon = \delta = .01, v_{\max} = 10$ )  
Regina:  $L = 0.4, U = 2.5$  (Tight Bounds)

	Owned Dwelling		Furniture\Equipment	
	RD	RP	RD	RP
Family I				
SMCS	-0.056	1.100	0.012	1.000
SM	-0.055	0.992	0.017	0.919
Family II				
MCS-r	-0.048	1.073	0.008	0.952
MDI-r	-0.045	1.087	0.012	0.965
GMDI	-0.047	1.077	0.009	1.006
	Ladies' Clothing		Men's Clothing	
Family I				
SMCS	0.024	0.917	0.038	0.808
SM	0.025	0.917	0.024	0.801
Family II				
MCS-r	0.020	0.904	0.012	0.922
MDI-r	0.025	0.888	0.012	0.922
GMDI	0.021	0.938	0.018	0.917

Note: During the jackknifing procedure, the SM method failed to converge in ten iterations for four pseudo-replicates (out of a total of 111).

Table 2c

Difference in Point Estimates and Precision Relative to Regression Estimator ( $\alpha = .67, \beta = .8, \eta = .9, \epsilon = \delta = .01, v_{\max} = 10$ )  
Saskatoon:  $L = 0.88, U = 1.12$

	Owned Dwelling		Furniture\Equipment	
	RD	RP	RD	RP
Family I				
SMCS	-0.001	1.001	-0.001	0.999
SM	-0.000	1.001	-0.000	0.999
Family II				
MCS-r	0.000	0.999	0.000	1.000
MDI-r	0.002	0.997	0.002	0.994
GMDI	-0.000	1.007	-0.000	0.990
	Ladies' Clothing		Men's Clothing	
Family I				
SMCS	0.000	1.013	-0.001	0.999
SM	-0.000	1.002	-0.000	0.998
Family II				
MCS-r	0.000	0.990	0.000	0.994
MDI-r	0.002	1.001	0.002	0.983
GMDI	0.000	0.977	-0.000	0.990

## Notes:

1. For the unrestricted (or no bounds) case, the corresponding measures for the raking (MDI-u) method relative to regression are (0.002, 1.000), (0.002, 1.000), (0.002, 1.002) and (0.002, 0.995) for the four study variables respectively.
2. During the jackknifing procedure, the SM method failed to converge in ten iterations for two pseudo-replicates (out of a total of 94).

convergence problems for [0.475, 2.11]. For Saskatoon (Table 1), under the chosen bounds each method takes only one or two iterations. With  $v_{\max}$  as 10, as bounds are tightened to [0.92, 1.08], SM does not converge. At [0.93, 1.07], SMCS, MCS-r, and MDI-r have convergence problems, and finally at [0.96, 1.06], GMDI has problems.

## 4. DISCUSSION

Although numerical results for a few variables for two different domains considered in this paper are quite limited to draw general conclusions, the results based on a descriptive analysis are nevertheless interesting and may provide some indications which might be useful in practice. These can be summarized in the following observations. For loose bounds, all the restricted methods seem to perform almost at par with the regression method. However, for tight bounds, there seem to be a difference in point estimates and especially in estimated precision. This observation clearly needs further study in light of the fact that all methods are asymptotically equivalent to the regression method. A simulation study in this regard would be desirable. The recent study of Stukel, Hidioglou, and Särndal (1996) sheds some light on this issue. Moreover, for tight bounds, there may not be convergence

under the specified number of iterations even if a solution exists. This problem may be more apparent in dealing with jackknife replicates. Therefore, caution should be exercised in choosing the maximum number of iterations for tight bounds. Finally, in practice, it is possible that even with minimal requirements on BCs and RRs, none of the calibration estimators converge within a reasonable number of iterations. In this situation, it would be of interest to investigate whether the (asymptotic) design consistency of calibration estimators could be preserved while allowing deviation from BCs. The idea of using ridge regression by Bardsley and Chambers (1984), although not in the design-based context, may be useful for this purpose. This problem is currently being investigated in collaboration with J.N.K. Rao.

## APPENDIX

Here we provide computational algorithms for all seven methods of weight adjustment. These algorithms were used to write computer programs in GAUSS software for the numerical examples presented in this paper.

In all the methods, some form of the following expression denoted by the  $n$ -vector  $f^{(v)}$ , is used repeatedly for computing  $c_k^{(v)}$  for  $v = 1, 2, \dots$

$$f^{(v)} \equiv X(X' \Gamma_{v-1} X)^{-1}(\tau_x - \hat{\tau}_x^{(v-1)}) \quad (1)$$

where  $\Gamma_{(v-1)}$  is an  $n \times n$  diagonal matrix defined below in the algorithm for each method. Initially  $\Gamma_0 = \text{diag}(h)$  and  $\hat{\tau}_x^{(0)} = \sum x_k h_k$ .

### A1. METHOD 1 (MCS-u)

The solution is non-iterative and is given in two steps as follows.

- (i) Compute  $f_k^{(1)}$ ,  $k = 1$  to  $n$  from (1) by setting  $\Gamma_{(v-1)} = \Gamma_0$ .
- (ii) Compute  $g_k$  as  $1 + f_k^{(1)}$  and then  $c_k^{\text{MCS-u}}$  as  $h_k g_k$ .

### A2. METHOD 2 (MDI-u)

The solution is obtained iteratively by the following steps for  $v = 1, 2, \dots$

- (i) Set the tolerance level  $\delta \geq 0$  for meeting BCs at some small value.
- (ii) For the  $v$ -th iteration, compute  $f_k^{(v)}$ ,  $k = 1$  to  $n$ , from (1) by setting  $\Gamma_{v-1} = \text{diag}(c_k^{(v-1)})$ .
- (iii) For  $v = 1, 2, \dots$  compute  $g_k^{(v)}$  as  $g_k^{(v-1)} \exp(f_k^{(v)})$ ,  $g_k^{(0)} = 1$  and then  $c_k^{(v)}$  from  $h_k g_k^{(v)}$ .
- (iv) Repeat steps (ii)-(iii) until the BCs are met up to the tolerance level  $\delta$  or the number of iterations is at its maximum,  $v_{\max}$ . The last iteration gives  $c_k^{\text{MDI-u}}$ .

### A3. METHOD 3 (SMCS)

The solution is obtained iteratively as follows.

- (i) Set the RRs, i.e., choose  $L$  and  $U$ ,  $L < 1 < U$ .
- (ii) Set the tolerance level  $\epsilon \geq 0$  at a small value for meeting the RRs.

- (iii) Choose a parameter  $\alpha$  between 0 and 1 (e.g. 2/3) and set  $L' = \alpha L + 1 - \alpha$ ,  $U' = \alpha U + 1 - \alpha$ . The default value of 1 for  $\alpha$  is also allowed in which case  $L' = L$ ,  $U' = U$ .
- (iv) For the  $v$ -th iteration with  $g_k^{(0)} = 1$ , define  $\xi_k^{(v-1)} = (g_k^{(v-1)} - 1)/(L' - 1)$  if  $g_k^{(v-1)} \leq 1$ ;  $(g_k^{(v-1)} - 1)/(U' - 1)$  otherwise.
- (v) Choose another parameter  $\beta$  between 0 and 1 (e.g., 4/5). Set  $q_k^{(v-1)} = 1$  if  $\xi_k^{(v-1)} < 1/2$ ;  $1 - \beta(\xi_k^{(v-1)} - 1/2)^2$  if  $1/2 \leq \xi_k^{(v-1)} < 1$ ;  $(1 - \beta/4)/\xi_k^{(v-1)}$  if  $\xi_k^{(v-1)} \geq 1$  and then define for  $v = 1, 2, \dots$ ,  $q_k^{[v-1]} = q_k^{(0)} \dots q_k^{(v-1)}$  where  $q_k^{(0)} = 1$ . Note compounding of  $q$ -factors in defining  $q_k^{[v-1]}$ .
- (vi) Compute  $f_k^{(v)}$  from (1) by setting  $\Gamma_{v-1} = \text{diag}(h_k q_k^{[v-1]})$ , and  $\hat{\tau}_x^{(v-1)} = \hat{\tau}_x^{(0)}$  for all  $v$ .
- (vii) Find  $g_k^{(v)}$  as  $1 + q_k^{[v-1]} f_k^{(v)}$  and then  $c_k^{(v)}$  as  $h_k g_k^{(v)}$ .
- (viii) Repeat steps (iv)-(vii) until the RRs are met up to the tolerance level  $\epsilon$  or  $v = v_{\max}$ . The last iteration gives  $c_k^{\text{SMCS}}$ . The value of  $\beta$  should remain the same at each iteration.

### A4. METHOD 4 (SM)

This method consists of the following steps performed iteratively.

- (i)-(ii) Same as in Method 3.
- (iii) Choose parameters  $\alpha, \eta$ ,  $0 < \alpha \leq \eta \leq 1$ , (e.g.,  $\alpha = 2/3$ ,  $\eta = 9/10$ ) and define
 
$$L' = \alpha L + (1 - \alpha), \quad U' = \alpha U + (1 - \alpha)$$

$$L'' = \eta L + (1 - \eta), \quad U'' = \eta U + (1 - \eta).$$

The default option for  $\alpha$  and  $\eta$  is 1 in which case  $L' = L'' = L$ ,  $U' = U'' = U$ .

- (iv) (Shrinkage). The  $c_k^{(v)}$  from the  $v$ -th iteration is shrunk to obtain  $c_k^{(v)*}$  according to  $c_k^{(v)*} = L' h_k$  if  $c_k^{(v)} < L' h_k$ ;  $U' h_k$  if  $c_k^{(v)} > U' h_k$ ;  $c_k^{(v)}$  otherwise. For  $v = 0$ ,  $c_k^{(0)} = c_k^{(0)*} = h_k$ .
- (v) (Minimization). Find  $f_k^{(v)}$  from (1) by setting  $\Gamma_{v-1} = \text{diag}(c_k^{(v-1)*})$  and  $\hat{\tau}_x^{(v-1)} = \hat{\tau}_x^{(v-1)*}$ .
- (vi) Compute  $g_k^{(v)}$  as  $g_k^{(v-1)*} (1 + f_k^{(v)})$  where  $g_k^{(v-1)*} = c_k^{(v-1)*}/h_k$  and then  $c_k^{(v)}$  from  $h_k g_k^{(v)}$ .
- (vii) Repeat steps (iv)-(vi) until the RRs are satisfied up to tolerance  $\epsilon$  or  $v = v_{\max}$ . The last iteration gives  $c_k^{\text{SM}}$ .

### A5. METHOD 5 (MCS-r)

The iterative algorithm consists of the following steps.

- (i) Set  $L$  and  $U$ .
- (ii) Set the tolerance level  $\delta \geq 0$  for meeting the BCs.
- (iii) Compute  $f_k^{(v)}$  from (1) by setting  $\Gamma_{v-1} = \text{diag}(h_k a_k^{(v-1)})$  where  $a_k^{(v-1)} = 1$  if  $g_k^{(v-1)}$  was truncated to  $L$  or  $U$ , and 0 otherwise.
- (iv) Set  $g_k^{(0)} = 1$  and compute  $g_k^{(v)}$  as  $g_k^{(v-1)} + f_k^{(v)}$  if  $L \leq g_k^{(v)} \leq U$ ; otherwise truncate  $g_k^{(v)}$  to  $L$  or  $U$  as the case may be, and then  $c_k^{(v)}$  as  $h_k g_k^{(v)}$ .
- (v) Repeat steps (iii)-(iv) until BCs are met at the tolerance level  $\delta$  or  $v = v_{\max}$ . The last iteration gives  $c_k^{\text{MCS-r}}$ .



**A6. METHOD 6 (MDI-r)**

The iterative algorithm consists of the following steps.

- (i)-(ii) Same as in Method 5.
- (iii) Compute  $f_k^{(v)}$  from (1) by setting  $\Gamma_{v-1} = \text{diag}(c_k^{(v-1)} a_k^{(v-1)})$  where  $a_k^{(v-1)}$  is defined as in Step (iii) of Method 5.
- (iv) Set  $g_k^{(0)} = 1$  and compute  $g_k^{(v)} = g_k^{(v-1)} \exp(f_k^{(v)})$  if  $L \leq g_k^{(v)} \leq U$ ; otherwise truncate  $g_k^{(v)}$  to  $L$  or  $U$  as the case may be, and then  $c_k^{(v)}$  as  $h_k g_k^{(v)}$ .
- (v) Repeat steps (iii)-(iv) until BCs are satisfied at tolerance  $\delta$  or  $v = v_{\max}$ . The last iteration gives  $c_k^{\text{MDI-r}}$ .

**A7. METHOD 7 (GMDI)**

The iterative algorithm consists of the following steps.

- (i)-(ii) Same as in Method 5.
- (iii) Compute  $f_k^{(v)}$  from (1) by setting  $\Gamma_{v-1} = \text{diag}(h_k d_k^{(v-1)})$  where  $d_k^{(v-1)}$  is analogous to  $d_k^{(1)}$  of Section 2.7.
- (iv) Using  $\mathbf{x}_k' \lambda^{(v)} = \mathbf{x}_k' \lambda^{(v-1)} + f_k^{(v)}$ , find  $g_k^{(v)}$  from the formula for  $g_k$  given in Section 2.7, and then  $c_k^{(v)}$  as  $h_k g_k^{(v)}$ .
- (v) Repeat steps (iii)-(iv) until BCs are met at tolerance  $\delta$  or  $v = v_{\max}$ . The last iteration gives  $c_k^{\text{GMDI}}$ .

**ACKNOWLEDGEMENTS**

The authors are grateful to C.-E. Särndal for helpful discussions and to U. Nevraumont for providing the FAMEX data. The authors are also grateful to referees for very useful comments. The first author's research was supported in part by an NSERC grant held at Carleton University, Ottawa.

**REFERENCES**

BARDSLEY, P., and CHAMBERS, R.L. (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics*, 33, 290-299.

DEVILLE, J.-C., and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

DEVILLE, J.-C., SÄRNDAL, C.-E., and SAUTORY, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.

HUANG, E.T., and FULLER, W.A. (1978). Nonnegative regression estimation for sample survey data. *Proceedings of the Social Statistics Section, American Statistical Association*, 300-305.

FULLER, W.A., LOUGHIN, M.M., and BAKER, H.D. (1994). Regression weighting in the presence of nonresponse with application to the 1987-88 Nationwide Food Consumption Survey. *Survey Methodology*, 20, 75-85.

LEMAÎTRE, G., and DUFOUR, J. (1987). An integrated method for weighting persons and families. *Survey Methodology*, 13, 199-207.

SÄRNDAL, C.-E. (1980). On  $\pi$ -inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 67, 639-650.

SINGH, A.C. (1993). On weight adjustment in survey sampling. Discussion paper for the 18th meeting of the Advisory Committee on Statistical Methods, Statistics Canada, Ottawa, October 25-26.

SINGH, A.C., and MOHL, C.A. (1997). Calibration estimators with application to FAMEX survey and computer program documentation. Methodology Branch Working Paper, Statistics Canada.

SINGH, M.P., DREW, J.D., GAMBINO, J.G., and MAYDA, F. (1990). *Methodology of the Canadian Labour Force Survey: 1984-1990*. Catalogue No. 71-526, Statistics Canada.

STUKEL, D.M., and BOYER, R. (1992). Calibration estimation: an application to the Canadian Labour Force Survey. Methodology Branch Working Paper, SSMD 92-009E, Statistics Canada.

STUKEL, D.M., HIDIROGLOU, M.A., and SÄRNDAL, C.-E. (1996). Variance estimation for calibration estimators: A comparison of jackknifing versus Taylor linearization. *Survey Methodology*, 22, 117-125.





# Variance Estimation for Calibration Estimators: A Comparison of Jackknifing Versus Taylor Linearization

DIANA M. STUKEL, MICHAEL A. HIDIROGLOU and CARL-ERIK SÄRNDAL<sup>1</sup>

## ABSTRACT

The use of auxiliary information in estimation procedures in complex surveys, such as Statistics Canada's Labour Force Survey, is becoming increasingly sophisticated. In the past, regression and raking ratio estimation were the commonly used procedures for incorporating auxiliary data into the estimation process. However, the weights associated with these estimators could be negative or highly positive. Recent theoretical developments by Deville and Särndal (1992) in the construction of "restricted" weights, which can be forced to be positive and upwardly bounded, has led us to study the properties of the resulting estimators. In this paper, we investigate the properties of a number of such weight generating procedures, as well as their corresponding estimated variances. In particular, two variance estimation procedures are investigated via a Monte Carlo simulation study based on Labour Force Survey data; they are Jackknifing and Taylor Linearization. The conclusion is that the bias of both the point estimators and the variance estimators is minimal, even under severe "restricting" of the final weights.

**KEY WORDS:** Auxiliary information; Raking ratio estimators; Regression estimators; Restricted weighting.

## 1. INTRODUCTION

Auxiliary information has many uses in survey sampling. One typical use is its incorporation at the estimation stage through the use of regression estimators or raking ratio estimators. For these estimators, a unit's sampling weight is multiplied by an adjustment factor to produce the final weight. A well-known shortcoming associated with the regression estimator is that some of the adjustment factors may be negative, resulting in negative final weights. On the other hand, for the raking ratio estimator, some adjustment factors may be very large and positive, resulting in unduly large final weights. These shortcomings can be overcome by considering a family of estimators, known as "calibration estimators". Developed by Deville and Särndal (1992), the estimators in this family incorporate auxiliary information, and in certain cases, non-negative weights can be ensured by prespecifying lower and upper bounds on the weights. These "calibration" weights are obtained by minimizing functions which measure the distances between original sampling weights and final calibrated weights, while respecting a set of benchmarking constraints. Huang and Fuller (1978) and Singh and Mohl (1996) have developed similar estimators which maintain the above properties. Ordinarily, there are very small differences between the point estimates corresponding to the various distance functions.

Historically, Statistics Canada's Labour Force Survey (LFS) has used, at different points in time, both the Taylor and Jackknife variance estimation techniques in tandem with regression and raking ratio estimators. Recently, the LFS has also allowed for the option of using other calibration estimators in addition to the previously available regression

estimator, to eliminate the problem of potential negative weights. It is therefore of interest to investigate the behaviour of these point estimators and their corresponding Taylor and Jackknife variance estimators, particularly for those estimators that allow bounding on the weights. Therein lies the main focus of this paper. Now, both the Taylor and the Jackknife have their advantages. The Taylor method is computationally much less intensive than the Jackknife method, but requires working out new expressions for each different parameter that is considered; this is particularly a burden in multipurpose surveys where many different parameters may be of interest. On the other hand, for the Jackknife method, cumbersome variance expressions need not be derived for each new parameter; only the functional form of the point estimator itself is required.

The paper is structured as follows: section 2 provides the theoretical underpinnings of calibration estimation and introduces a family of related distance functions. In section 3, variances for calibration estimators are discussed. Section 4 provides the results of a Monte Carlo simulation study, in which the bias of both the point estimators and their corresponding Taylor and Jackknife variance estimators (relative to a "true" variance) is tracked, for a variety of distance functions from calibration theory. In section 5, some concluding remarks are made.

## 2. DISTANCE FUNCTIONS AND CALIBRATION ESTIMATORS

We begin by introducing the basic idea behind calibration estimation. Let  $U = \{1, \dots, k, \dots, N\}$  denote the index set for

<sup>1</sup> Diana M. Stukel, Household Survey Methods Division, and Michael A. Hidiroglou, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6; Carl-Erik Särndal, Département de Mathématiques et de Statistique, Université de Montréal, C.P. 6128, Succursale A, Montréal, P.Q., H3C 3J7.

the  $N$  units of a finite population of units. In survey sampling, one is often interested in estimating parameters of a finite population such as totals, means and ratios. For the sake of simplicity, we will focus on totals, although the ideas presented in this paper may easily be extended to include other parameters. Thus, suppose the objective is to estimate the population total  $Y = \sum_{k \in U} y_k$ , where  $y_k$  is the value of  $y$ , the variable of interest for the  $k$ -th population unit.

A probability sample  $s$  is drawn from  $U$  by a given sampling design which induces the inclusion probabilities  $\pi_k = P(k \in s)$ . These are assumed known and positive. Let  $a_k = 1/\pi_k$  be the sampling weight associated with the  $k$ -th unit. Finally, let the auxiliary information be specified in the form of known population totals of one or more auxiliary variables.

An elementary estimator of  $Y$  is the Horvitz-Thompson (HT) estimator:

$$\hat{Y}_a = \sum_{k \in s} a_k y_k.$$

The HT estimator possibly but not necessarily (depending on the sampling design) incorporates auxiliary information at the design stage only; what is sought is an improved estimator which incorporates the auxiliary information at the estimation stage, as well. The incorporation of auxiliary information can be reflected in the creation of new weights, denoted by  $w_k$ ;  $k \in s$ . The new estimator is then of the form:

$$\hat{Y}_w = \sum_{k \in s} w_k y_k. \quad (2.1)$$

The approach of Deville and Särndal (1992) and Deville, Särndal and Sautory (1993) involves determining these new weights  $\{w_k; k \in s\}$  by making them as close as possible to the original sampling weights  $\{a_k; k \in s\}$  according to a specified distance function. Constraints placed on the new weights are such that, when applied to each of the auxiliary variables, the known population total  $X$  is reproduced. That is,

$$\sum_{k \in s} w_k \mathbf{x}_k = \mathbf{X} \quad (2.2)$$

is required to hold, leading to a problem in constrained minimization. Here  $\mathbf{x}_k' = (x_{1k}, x_{2k}, \dots, x_{pk})$  is a vector of length  $p$  containing the values of the auxiliary variables for the  $k$ -th individual, and the auxiliary information available from an external source is summarized by the known vector total  $\mathbf{X} = \sum_{k \in U} \mathbf{x}_k$ .

We denote the distance from  $w_k$  to  $a_k$  by  $F^*(w_k, a_k)$ . Deville and Särndal (1992) limit their discussions to distance functions of the form  $F^*(w_k, a_k) = a_k c_k F(w_k/a_k)$  where  $w_k/a_k = g_k$ , the ratio of the final calibrated weight to original sampling weight, is called the "g-factor". Here  $c_k$  is a known positive weight unrelated to  $a_k$ ; the uniform weighting  $c_k = 1$  is often used in applications. Note that equation (2.1) can alternatively be written as:

$$\hat{Y}_w = \sum_{k \in s} a_k g_k y_k.$$

It is assumed that  $F$  is non-negative and convex, and that  $F(1) = 0$ , implying that when  $w_k = a_k$  the distance between the weights is zero. Moreover, it is required that  $F'$  is continuous, one-to-one, and that  $F'(1) = 0$  and  $F''(1) > 0$  which makes  $w_k = a_k$  a local minimum. (See Deville, Särndal and Sautory 1993.) The total distance,  $\sum_{k \in s} a_k c_k F(w_k/a_k)$ , is minimized subject to the constraint (2.2). That is,

$$\sum_{k \in s} a_k c_k F(w_k/a_k) - \lambda' \left( \sum_{k \in s} w_k \mathbf{x}_k - \mathbf{X} \right)$$

is minimized with respect to the  $w_k$ , where  $\lambda$  is a  $p$ -vector of Lagrange multipliers. Differentiating with respect to  $w_k$ , equating to zero, and solving for  $w_k$  leads to the calibrated weights  $w_k = a_k g_k = a_k g(\lambda' \mathbf{x}_k / c_k)$  where  $g$  is the inverse function of  $f$  and  $f(z) = dF(z)/dz$ . To compute  $w_k$ , one must first obtain  $\lambda$  as the solution of the calibration equation implied by (2.2), namely,

$$\sum_{k \in s} a_k g(\lambda' \mathbf{x}_k / c_k) \mathbf{x}_k = \mathbf{X}. \quad (2.3)$$

The solution of this (possibly) nonlinear system of  $p$  equations in  $p$  unknowns may require the use of some iterative procedure, such as the Newton-Raphson method.

A number of distance functions are considered by Deville and Särndal (1992), Huang and Fuller (1978) and Singh and Mohl (1996). Two important distance functions which we first discuss are the Generalized Least Squares (GLS) distance function and the Raking Ratio (RR) distance function, both given in Deville and Särndal (1992).

The GLS distance function is defined by:

$$\begin{aligned} F^*(w_k, a_k) &= F_{\text{GLS}}^*(w_k, a_k) \\ &= c_k (w_k - a_k)^2 / a_k = a_k c_k (w_k/a_k - 1)^2. \end{aligned} \quad (2.4)$$

It generates the well-known generalized regression estimator (GREG), which encompasses as special cases the ratio estimator, the simple regression estimator, and the simple post-stratified estimator, among others. It follows from (2.3) that the calibrated weights corresponding to the GLS distance function are:

$$w_k = a_k g_k = a_k [1 + (\mathbf{X} - \hat{\mathbf{X}}_a)' \left( \sum_{j \in s} a_j \mathbf{x}_j \mathbf{x}_j' / c_j \right)^{-1} \mathbf{x}_k / c_k]$$

where  $\hat{\mathbf{X}}_a = \sum_{k \in s} a_k \mathbf{x}_k$  is the HT estimator of  $\mathbf{X}$ . The corresponding estimator of  $Y$  can be written in the usual regression estimator form as

$$\hat{Y}_{w(\text{GREG})} = \hat{Y}_a + (\mathbf{X} - \hat{\mathbf{X}}_a)' \hat{\beta} \quad (2.5)$$

where

$$\hat{\beta} = \left( \sum_{k \in s} a_k \mathbf{x}_k \mathbf{x}_k' / c_k \right)^{-1} \sum_{k \in s} a_k \mathbf{x}_k y_k / c_k. \quad (2.6)$$



Thus, the regression estimator can be thought of as the HT estimator plus an adjustment term. A drawback of the GLS distance function is that it may give rise to negative weights, particularly if the system is overconstrained. In practice, negative weights are rare; however, it is desirable to eliminate them entirely since it may be difficult to give them any meaningful interpretation.

The Raking Ratio (RR) distance function is defined by:

$$\begin{aligned} F^*(w_k, a_k) &= F_{RR}^*(w_k, a_k) \\ &= c_k [w_k \log(w_k/a_k) - w_k + a_k] \\ &= a_k c_k [(w_k/a_k) \log(w_k/a_k) - (w_k/a_k) + 1]. \end{aligned} \quad (2.7)$$

Solving for  $g$ -factors using the RR distance function and the constraint defined by equation (2.3) can be shown to be equivalent to using the Iterative Proportional Fitting (IPF) algorithm of Deming and Stephan (1940) when calibrating on known marginals of frequency tables of dimension two or higher. Unlike the GLS distance function, which has a closed form solution, the calibration equations for the RR distance function can only be solved iteratively. Computer software exists for this purpose; for example, the CALMAR software (see Deville, Särndal and Sautory 1993) solves the calibration equations for the RR distance function using the Newton-Raphson method, rather than the IPF algorithm originally proposed by Deming and Stephan. The RR distance function always ensures positive weights; however, it also has the undesirable property that some of the resulting calibration weights can be excessively large.

Neither the possibility of negative weights produced by the GLS distance function nor the possibility of large positive weights produced by the RR distance function are desirable. One can define restricted distance functions whereby the range of the resulting weights  $w_k$  are limited. This is achieved by imposing restrictions on the distance function  $F(w_k/a_k)$  in such a way that the  $g$ -factors  $g_k = w_k/a_k$  are bounded within a prespecified interval. To this end, one can specify a lower bound  $L$  and an upper bound  $U$ , such that  $L < 1 < U$ . To guarantee positive weights, one would choose  $L > 0$ . Now, Deville and Särndal (1992) define restricted versions of the two distance functions given above; they are: the Restricted GLS (RGLS) distance function and the Restricted Raking Ratio (RRR) or Logit distance function. Two other methods of restricting final weights are proposed by Huang and Fuller (1978) and Singh and Mohl (1996). All four restricted distance functions are considered in this paper; they are also discussed in detail in Singh and Mohl (1996), but from a different perspective.

The Restricted GLS distance function is defined by:

$$F^*(w_k, a_k) = \begin{cases} c_k (w_k - a_k)^2 / a_k & \text{if } L < w_k/a_k < U \\ \infty & \text{otherwise.} \end{cases} \quad (2.8)$$

The Restricted RR (or Logit) distance function is defined by:

$$F^*(w_k, a_k) = F_{RRR}^*(w_k, a_k) = \begin{cases} A^{-1} c_k [(w_k/a_k - L) \log[(w_k/a_k - L)/(1 - L)] \\ \quad + (U - w_k/a_k) \log[(U - w_k/a_k)/(U - 1)]] & \text{if } L < w_k/a_k < U \\ \infty & \text{otherwise} \end{cases} \quad (2.9)$$

where  $A = (U - L)/\{(1 - L)(U - 1)\}$ . The specification  $L = 0$ ,  $U = \infty$  gives the RR distance function. It is easy to show that the Restricted GLS and Restricted RR distance functions share the property that the corresponding weights  $w_k$  satisfy  $L < w_k/a_k < U$ .

Now, Huang and Fuller (1978) propose a method for adjusting regression weights such that the calibration constraints given by equation (2.2) are satisfied and such that the  $g$ -factors are restricted to lie close to one. Singh and Mohl (1996) show that their method can be written in terms of minimizing a distance function which changes from iteration to iteration. Singh and Mohl also modify the original method to allow for arbitrary restrictions on the  $g$ -factors, similar to the restricted distance functions above, and show that the estimator resulting from the modified distance function is asymptotically equivalent to the regression estimator. The Modified Huang-Fuller (MHF) distance function is given by:

$$\begin{aligned} F^*(w_k^{(v-1)}, a_k) &= F_{MHF}^{(v)*}(w_k^{(v-1)}, a_k) \\ &= (w_k^{(v-1)} - a_k)^2 / a_k q_k^{(v-1)*}; \quad v = 1, 2, \dots \end{aligned} \quad (2.10)$$

where  $q_k^{(v-1)*} = q_k^{(v-1)} \dots q_k^{(1)} q_k^{(0)}$  with  $q_k^{(0)} = 1$  and where  $v$  is the iteration number. Here,

$$q_k^{(v-1)} = \begin{cases} 1 & \text{if } \xi_k^{(v-1)} < .5 \\ 1 - \delta (\xi_k^{(v-1)} - .5)^2 & \text{if } .5 \leq \xi_k^{(v-1)} < 1 \\ (1 - \delta/4) / \xi_k^{(v-1)} & \text{if } \xi_k^{(v-1)} \geq 1 \end{cases}$$

for  $\delta$  arbitrarily chosen such that  $0 < \delta < 1$ . Also

$$\xi_k^{(v-1)} = \begin{cases} (g_k^{(v-1)} - 1)/(L' - 1) & \text{if } g_k^{(v-1)} \leq 1 \\ (g_k^{(v-1)} - 1)/(U' - 1) & \text{otherwise} \end{cases}$$

where  $L' = \alpha L + 1 - \alpha$  and  $U' = \alpha U + 1 - \alpha$  for  $\alpha$  arbitrarily chosen such that  $0 < \alpha < 1$  and  $L$  and  $U$  are as in earlier restricted distance functions. The parameters  $\alpha$  and  $\delta$  serve to speed up the convergence of the iterative algorithm used to provide a solution. Singh and Mohl (1996) empirically test a variety of values for these parameters using large data sets, and suggest that  $\alpha = .67$  and  $\delta = .8$  work well in practice. Finally, the  $g$ -factor at each iteration is

$$g_k^{(v-1)} = \frac{1}{1 + (X - \hat{X}_w^{(v-2)})' \left( \sum_{j \in S} a_j q_j^{(v-2)*} x_j x_j' \right)^{-1} x_k}; \quad v = 2, 3, \dots$$

where  $\hat{X}_w^{(v-2)} = \sum_{k \in S} w_k^{(v-2)} x_k$ ;  $v = 2, 3, \dots$  and where  $w_k^{(v-2)} = a_k g_k^{(v-2)}$ ;  $v = 2, 3, \dots$ . Starting values are given by  $g_k^{(0)} = 1$  and  $w_k^{(0)} = a_k$ .

Singh and Mohl (1996) also propose a new distance function which changes from iteration to iteration called the Shrinkage-Minimization (SM) distance function, and show that the estimator resulting from this distance function is also asymptotically equivalent to the regression estimator. It is given by:

$$F^*(w_k^{(v-1)}, a_k) = F_{SM}^{(v)*}(w_k^{(v-1)}, a_k) \\ = (w_k^{(v-1)} - a_k^{(v-1)*})^2 / a_k^{(v-1)*}; \quad v = 1, 2, \dots \quad (2.11)$$

where

$$a_k^{(v-1)*} = \begin{cases} L' a_k & \text{if } w_k^{(v-1)} < L' a_k \\ U' a_k & \text{if } w_k^{(v-1)} > U' a_k \\ w_k^{(v-1)} & \text{otherwise.} \end{cases} \quad v = 2, 3, \dots$$

Terms in the above equations are defined as follows:  $L' = \alpha L + (1 - \alpha)$ ,  $U' = \alpha U + (1 - \alpha)$ ,  $L'' = \eta L + (1 - \eta)$  and  $U'' = \eta U + (1 - \eta)$  for  $\alpha$  and  $\eta$  arbitrarily chosen such that  $0 < \alpha < \eta \leq 1$ . As before, the parameters  $\alpha$  and  $\eta$  serve to speed up the convergence of the iterative algorithm used to provide a solution; Singh and Mohl (1996) suggest that  $\alpha = .67$  and  $\eta = .9$  work well in practice. Finally,  $w_k^{(v-1)} = a_k g_k^{(v-1)}$ ;  $v = 2, 3, \dots$  where

$$g_k^{(v-1)} = \frac{a_k^{(v-2)*}}{a_k} \left[ 1 + (X - \hat{X}_w^{(v-2)})' \left( \sum_{j \in S} a_j^{(v-2)*} x_j x_j' \right)^{-1} x_k \right]; \\ v = 2, 3, \dots$$

and where  $\hat{X}_w^{(v-2)}$  is as before. Starting values are given by  $a_k^{(0)*} = a_k$  and  $w_k^{(0)} = a_k$ .

A property of the Modified Huang-Fuller and Shrinkage-Minimization distance functions is that the calibration constraints (equation (2.2)) are met at every iteration whereas the range restrictions on the  $g$ -factors are met only upon convergence. For the Restricted GLS and Restricted Raking Ratio distance functions, the range restrictions on the  $g$ -factors are met at every iteration whereas the calibration constraints are only met upon convergence. Now, it is often useful to specify an upper bound on the number of iterations to convergence; this feature may be programmed into the iterative algorithm for operational expediency. If this upper bound is exceeded due to slow convergence, the iterative algorithm may be terminated prematurely. Regardless, for the Modified Huang-Fuller and Shrinkage-Minimization distance functions, the calibration constraints will be met. Likewise,

for the Restricted GLS and Restricted Raking Ratio distance functions, the range restrictions will be met.

Now, the behaviour of the  $g$ -factors from some of the distance functions has been studied extensively; see, for example, Deville, Särndal and Sautory (1993). Stukel and Boyer (1992) empirically show that the GLS and RR distance functions, as well as their restricted counterparts having loose bounds imposed on them, give  $g$ -factors whose distributions over a given data set adhere to normality rather closely. However, as the bounds on the restricted distance functions are squeezed together more closely, the distributions exhibit a “pile-up” of  $g$ -factors at the lower and upper bounds. Regardless, even under extreme squeezing, the restricted distance functions seem to give point estimates that are close to their unrestricted counterparts, as the results of our empirical study will verify. However, the biases of both the point and variance estimators under extreme squeezing on the restricted distance functions have not been investigated. This investigation is of interest to surveys such as the LFS, where an augmentation to the current estimation system has been implemented, which now allows users the option of choosing from amongst the Restricted GLS distance function and the Shrinkage-Minimization distance function, in addition to the previously available GLS distance function.

### 3. VARIANCE ESTIMATION FOR CALIBRATION ESTIMATORS

The exact variance of the calibration estimator  $\hat{Y}_w$  is intractable since the point estimator itself is nonlinear. In addition, there is no explicit unbiased method of variance estimation. Therefore, approximately unbiased methods, such as the Taylor and the Jackknife, are often used in practice.

Now, for stratified multistage designs, “with replacement” sampling is not often used in practice since the possibility of drawing the same unit more than once is unappealing. Therefore, the preponderance of surveys use “without replacement” sampling, at least at the first stage of sampling. Even so, if the first stage sampling fraction is small (say, less than 10 percent as a rule of thumb), it may be reasonable to use a simplified variance formula that assumes “with replacement” sampling at the first stage of sampling. For the generalized regression estimator (GLS distance function) under a stratified multistage design this simplification of the variance estimator yields:

$$\hat{V}_T^*(\hat{Y}_{w(\text{GREG})}) = \sum_{h=1}^L \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \left[ \sum_{k \in s_{hi}} a_{hik} e_{hik} - \frac{1}{n_h} \sum_{i=1}^{n_h} \sum_{k \in s_{hi}} a_{hik} e_{hik} \right]^2 \quad (3.1)$$

where  $s_{hi}$  is the sample of individuals in the  $i$ -th primary sampling unit (PSU) and the  $h$ -th stratum,  $a_{hik}$  is the original sampling weight under the stratified multi-stage design for



sampled individual  $k$  in PSU  $i$  and stratum  $h$ , and  $n_h$  is the number of sampled PSUs in stratum  $h$ . Also  $e_{hik} = y_{hik} - \mathbf{x}'_{hik} \hat{\beta}$  is the estimated residual associated with the regression estimator where  $\hat{\beta} = (\sum_{hik \in s} a_{hik} \mathbf{x}_{hik} \mathbf{x}'_{hik} / c_{hik})^{-1} \sum_{hik \in s} a_{hik} \mathbf{x}_{hik} y_{hik} / c_{hik}$ . For many designs, the "with replacement" formula given by (3.1) overestimates the true variance (see Särndal, Swensson and Wretman 1992, section 4.6). Note that although, technically speaking, this simplified variance estimator is *not* the Taylor variance estimator, it is often referred to as such for historical reasons and so will it be in this paper.

An improvement to equation (3.1), which includes the  $g$ -factor in the variance formula (recall that  $w_{hik} = a_{hik} g_{hik}$ ), is suggested by Hidiroglou, Fuller and Hickman (1980). It is given by:

$$\hat{V}_T(\hat{Y}_{w(\text{GREG})}) = \sum_{h=1}^L \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \left[ \sum_{k \in s_{hi}} w_{hik} e_{hik} - \frac{1}{n_h} \sum_{i=1}^{n_h} \sum_{k \in s_{hi}} w_{hik} e_{hik} \right]^2 \quad (3.2)$$

An analogue of equation (3.2) is also suggested by Särndal (1982) in the context of two-stage sampling, but for Yates-Grundy type variance estimators. Now Deville and Särndal (1992) show that any distance function which obeys a set of general conditions will produce an estimator that is asymptotically equivalent to the one produced by the GLS distance function, that is,  $\hat{Y}_{w(\text{GREG})}$  given by (2.5). Singh and Mohl (1996) extend this result to include the Modified Huang-Fuller and Shrinkage-Minimization distance functions. As a result, the asymptotic variance of the calibration estimator  $\hat{Y}_w$  can be considered to be roughly equal to that of  $\hat{Y}_{w(\text{GREG})}$ . This observation leads to a method for estimating the Taylor variance which is common to all calibration estimators, namely, to estimate the variance of  $\hat{Y}_w$  using a modification of the Taylor variance estimator employed for  $\hat{Y}_{w(\text{GREG})}$ , rather than rederiving the Taylor formula for each of the distance functions separately. Thus, whenever a variance estimator associated with a distance function different from the GLS is required, equation (3.2) is used, replacing the final weights  $\{w_{hik}\}$  from the GLS distance function with those from the distance function in question.

It is straightforward to apply the Jackknife procedure to obtain a variance estimator for  $\hat{Y}_w$ , regardless of the distance function used to obtain the final calibrated weights. An expression for the variance formula under a stratified multi-stage design using with replacement sampling at the first stage is given by:

$$\hat{V}_J(\hat{Y}_w) = \sum_{h=1}^L \frac{n_h - 1}{n_h} \sum_{i=1}^{n_h} (\hat{Y}_w(hi) - \hat{Y}_w)^2 \quad (3.3)$$

where  $\hat{Y}_w(hi)$  is often referred to as the "replicate estimator"; "replicates" are formed by taking what remains of the sample after removing PSU  $i$  from stratum  $h$ . Thus,  $\hat{Y}_w(hi)$  is calculated by recomputing  $\hat{Y}_w$  after removing the  $i$ -th PSU

from the  $h$ -th stratum,  $h = 1, \dots, L$ ;  $i = 1, \dots, n_h$ , i.e., with the original sampling weights altered to reflect the PSU removal and the  $g$ -factors recalculated based on the reduced sample or replicate. Finally, the Jackknife estimator is constructed by repeatedly removing PSUs one at a time, calculating the corresponding replicate estimator, and then assembling the final estimator using (3.3). The Jackknife variance estimator given by (3.3) is the most conservative among the four variations suggested in the extensive discussion on the subject by Wolter (1985).

It is interesting to note that, for the GREG estimator, Yung and Rao (1996) obtain (3.2) as an approximation to the Jackknife variance estimator given by (3.3); they call (3.2) the "Jackknife Linearization Variance Estimator". Their simulation study shows that biases (both conditional and unconditional) of the Taylor variance estimator (equation (3.1)), the Jackknife Linearization variance estimator (equation (3.2)) and the Jackknife variance estimator (equation (3.3)) behave similarly. While their simulation focuses on variance estimators for the unrestricted GREG estimator, our simulation study, which we discuss next, focuses on variance estimators for the GREG as well as for estimators based on other restricted and unrestricted distance functions.

## 4. MONTE CARLO SIMULATION STUDY

### 4.1 Design of the Study

In order to compare the performance of the calibration estimators and their corresponding Taylor and Jackknife variance estimators, we undertook a Monte Carlo simulation study, in which we investigated their finite sample design-based frequentist properties.

December 1990 Labour Force Survey (LFS) sample data for the province of Newfoundland was used to simulate a finite population, from which repeated samples were drawn. The LFS is the largest ongoing household sample survey conducted by Statistics Canada. Monthly data relating to the labour market is collected using a complex multi-stage sampling design with several levels of stratification. The details of the design of the survey prior to the 1991 redesign can be found in Singh, Drew, Gambino and Mayda (1990). In general, provinces are stratified into "economic regions", which are large areas of similar economic structure; Newfoundland has four such economic regions. The economic regions are further substratified into "self-representing units" (SRUs) and "non self-representing units" (NSRUs), which are, in turn, further substratified into lower level substrata. SRUs are cities whose population exceeds 15,000, such as St. John's and Cornerbrook, in the case of Newfoundland. Now, the lowest level of stratification in Newfoundland yielded 45 strata, each of which contained less than 6 primary sampling units (PSUs), which was an insufficient number from which to sample, for the purposes of the simulation. Thus, the 45 strata were collapsed down to 18, each containing between 6 and 18 PSUs. In collapsing the strata,

economic regions were kept intact, as were the Census Metropolitan Areas (CMAs) of St. John's and Cornerbrook.

For the Monte Carlo study,  $R = 4,000$  samples, each of size approximately 1,000, were drawn from the Newfoundland "population" (which was of size 9,152), according to a two-stage design. For collapsed strata belonging to NSRUs, two PSUs were selected at the first stage using Probability Proportional to Size (PPS) with replacement (WR) sampling, where the size measure used was the number of dwellings in the PSU. At the second stage, one in five dwellings were selected from the sampled PSUs using Simple Random Sampling (SRS) without replacement (WOR). For collapsed strata belonging to SRUs, three PSUs were selected at the first stage using PPS WR sampling. At the second stage, all the dwellings in the sampled PSUs were selected, reducing this part of the design to one-stage take-all cluster sampling. This feature was necessary since there were not enough dwellings per PSU to subsample in SRUs. The selection of two PSUs in NSRU strata versus three in SRU strata was driven by the fact that, in general, NSRU strata had fewer population PSUs from which to sample than did SRU strata. In all, there were 47 sampled PSUs. In either case (NSRUs or SRUs), all dwelling members were included in the sample. Although this design is a hybrid between a one and two-stage design, we shall refer to it as a two-stage design, for convenience.

We took  $Y$ , the total number of unemployed, to be the parameter of interest. This was calculated from the finite population by:  $Y = \sum_{k \in U} y_k = \sum_{k=1}^{9152} y_k$  where  $y_k = 1$  if individual  $k$  was unemployed; 0 otherwise. For each of the  $R = 4,000$  samples, we calculated  $\hat{Y}_w$ , the estimated total number of unemployed as  $\hat{Y}_w = \sum_{k \in s} w_k y_k$ . The  $\{w_k : k \in s\}$  were determined by the following six distance functions discussed earlier:

- (1) the Generalized Least Squares (GLS) Distance Function (equation (2.4)),
- (2) the Raking Ratio (RR) Distance Function (equation (2.7)),
- (3) the Restricted GLS (RGLS) Distance Function (equation (2.8)),
- (4) the Restricted RR (RRR) or Logit Distance Function (equation (2.9)),
- (5) the Modified Huang-Fuller (MHF) Distance Function ( $\alpha = .67$ ,  $\delta = .8$ ) (equation (2.10)), and
- (6) the Shrinkage-Minimization (SM) Distance Function ( $\alpha = .67$ ,  $\eta = .9$ ) (equation (2.11)).

For the latter four distance functions, the following four sets of bounds were imposed on each to restrict the minimization: (i)  $L = 0$ ,  $U = 4$ , (ii)  $L = .4$ ,  $U = 2$ , (iii)  $L = .68$ ,  $U = 1.6$  and (iv)  $L = .8$ ,  $U = 1.3$ . This yielded a total of eighteen point estimators. For each of the eighteen point estimators, the calibration used auxiliary information based on Census projections at the province level for 10 mutually exclusive and exhaustive age/sex categories (age categories:  $< = 14$ , 15-24, 25-44, 45-64,  $> = 65$  crossed with the two sexes) and the four economic regions of Newfoundland.

Thus, the auxiliary information for each individual was a vector of length fourteen having exactly two ones and twelve zeros. However, for computational purposes, the dimensionality of the vector had to be reduced to thirteen when using the Newton-Raphson procedure to solve equation (2.3). For the first four distance functions, we set  $c_k = 1$ .

For each of the  $R = 4,000$  samples and each of the eighteen point estimators, we calculated the Jackknife variance estimator given by equation (3.3). We also calculated the Taylor variance estimator given by equation (3.2), and the modification suggested in section 3 was used for distance functions other than the GLS. Note that since PPSWR, rather than PPSWOR, was used at the first stage of sampling, the use of the variance estimator given by equation (3.2) was entirely appropriate for our simulation. Finally, for the GLS distance function only, the formula (3.1) was calculated to observe the impact of omitting  $g$ -factors from the variance estimator.

For each of the six distance functions given above, a number of frequentist properties were investigated. These are given below.

(A) The Percent Relative Bias of the Estimated Number of Unemployed (with respect to the population value) is estimated by:

$$\frac{E_M(\hat{Y}_w) - Y}{Y} * 100 \quad (4.1)$$

where

$$E_M(\hat{Y}_w) = \frac{1}{R} \sum_{r=1}^R \hat{Y}_{w_r}$$

is the Monte Carlo expectation of the point estimator  $\hat{Y}_w$  taken over the  $R$  samples, and  $\hat{Y}_{w_r}$  is the value of  $\hat{Y}_w$  for sample  $r$ .

(B) The Percent Relative Bias of the Taylor/Jackknife Variance Estimator (with respect to the true variance) is estimated by:

$$\frac{(E_M(\hat{V}(\hat{Y}_w)) - V_{\text{true}})}{V_{\text{true}}} * 100 \quad (4.2)$$

where

$$E_M(\hat{V}(\hat{Y}_w)) = \frac{1}{R} \sum_{r=1}^R \hat{V}_r(\hat{Y}_w)$$

and

$$V_{\text{true}} = \frac{1}{R} \sum_{r=1}^R (\hat{Y}_{w_r} - E_M(\hat{Y}_w))^2$$

and  $\hat{V}_r(\hat{Y}_w)$  is the value of  $\hat{V}(\hat{Y}_w)$  (Taylor or Jackknife) for sample  $r$ .

(C) The Percent Coefficient of Variation of the Taylor/Jackknife Variance Estimator (with respect to the true variance) is estimated by:



$$\sqrt{\frac{\frac{1}{R} \sum_{r=1}^R (\hat{V}_r(\hat{Y}_w) - V_{\text{true}})^2}{V_{\text{true}}}} * 100 \quad (4.3)$$

i.e., the root mean squared error of the variance estimator divided by the true variance, expressed as a percentage. Although most studies focus on the *bias* of the variance estimators, it is also of secondary interest to look at the *coefficient of variation* of the variance estimators to see how variable the variance estimates themselves are.

Note that in equations (4.2) and (4.3), it may have been more appropriate to make comparisons relative to a "true mean squared error" rather than a "true variance". However, for our simulation, the relative biases were so small that the differences between the two types of comparisons are virtually negligible.

Finally, in order to assess the appropriateness of the choice of number of repeated samples, we calculated Monte Carlo errors, using as a measure the Percent Coefficient of Variation of  $E_M(\hat{V}(\hat{Y}_w))$ , given by:

$$\sqrt{\frac{\frac{1}{R^2} \sum_{r=1}^R [\hat{V}_r(\hat{Y}_w) - E_M(\hat{V}(\hat{Y}_w))]^2}{E_M(\hat{V}(\hat{Y}_w))}} * 100. \quad (4.4)$$

The Monte Carlo errors were found to be consistently low (between .99% and 3.60%) for both the Jackknife and Taylor using  $R = 4,000$ , indicating stable results.

## 4.2 Results of the Study

Table 1 gives the Percent Relative Bias of the Point Estimators (equation (4.1)) as well as the Percent Relative Bias of the Taylor and Jackknife Variance Estimators (equation (4.2)) and the Percent CVs of the Taylor and Jackknife Variance Estimators (equation (4.3)). The percent relative bias for all the point estimates (column two) is negligible, ranging in value from 0.10% to 0.52%, but much less than 1% in all cases. The fact that all point estimates have a similar bias seems reasonable, given the asymptotic equivalence of all calibration estimators to the regression estimator.

The third column gives the percent relative bias of the Taylor variance estimator. Here, the true variance is always underestimated, but never by more than 6.2%. In the case of the regression estimator, it appears to make little difference whether or not the  $g$ -factor is included in the variance formula (equation (3.1) versus (3.2)); the bias improves only slightly for the case of the  $g$ -factor included (-5.82% versus -6.01%). The Jackknife variance estimator (column four), on the other hand, outperforms the Taylor variance estimator uniformly. The Jackknife almost always underestimates the true variance, but by less than 2% in all cases.

To produce a solution, all distance functions but the GLS required an iterative algorithm. This being the case, some of the 4,000 samples experienced convergence problems, particularly in the case of extreme bounding on the  $g$ -factors. Those samples for which the algorithm did not converge were discarded. Thus, they did not contribute to the various Monte Carlo measures. The number of such discarded samples is

**Table 1**  
Percent Relative Bias of the Point Estimators, and Percent Relative Bias and Percent CV of the Taylor and Jackknife Variance Estimators (Sample Size About 1000)

Distance Function	Percent Relative Bias Point Estimator	Percent Relative Bias Taylor Variance	Percent Relative Bias Jackknife Variance	Percent CV Taylor Variance	Percent CV Jackknife Variance	Number of Discarded Samples (From 4000)
GLS (Regression)	.11	-6.01 (eq 3.1) -5.82 (eq 3.2)	-1.73	60.79 (eq 3.1) 59.60 (eq 3.2)	62.86	0
Restricted GLS ( $L = 0, U = 4$ ) ( $L = .4, U = 2$ )	.11	-5.82	-1.73	59.60	62.86	0
	.10	-5.36	-1.27	59.93	63.21	32
Raking Ratio	.52	-6.20	0.84	59.45	63.35	0
Restricted RR ( $L = 0, U = 4$ ) ( $L = .4, U = 2$ )	.50	-6.09	-0.31	59.48	63.47	0
	.46	-5.69	-0.39	59.81	64.21	32
Modified Huang-Fuller ( $L = 0, U = 4$ ) ( $L = .4, U = 2$ )	.11	-5.82	-1.73	59.60	62.86	0
	.10	-5.36	-1.20	59.94	63.27	32
Shrinkage-Minimization ( $L = 0, U = 4$ ) ( $L = .4, U = 2$ )	.11	-5.82	-1.73	59.60	62.86	0
	.10	-5.36	-1.27	59.94	63.25	32

indicated in the last column of Table 1. In the case of extreme bounds ( $L = .68$ ,  $U = 1.6$  and  $L = .8$ ,  $U = 1.3$ ), so many samples were discarded (between 231 and 234 for the cases  $L = .68$ ,  $U = 1.6$  and between 1,562 and 1,602 for the cases  $L = .8$ ,  $U = 1.3$ ) that the results were not considered reliable, and so are not reported here. However, these tighter bounds were of interest, so the simulation was rerun using approximately double the sample size (increase from roughly 1,000 to 2,000). Note that Deville and Särndal (1992) show that convergence is achieved for all distance functions with probability one as the sample size increases.

Columns five and six of Table 1 give the Percent CVs of the Taylor and Jackknife Variance Estimators. The coefficients of variation are similar for all distance functions, ranging in value from 59.45% to 64.21%. However, the CVs corresponding to the Jackknife are always slightly larger than that of Taylor. Coefficients of variation of this magnitude, although large, have been encountered in other simulation studies relating to variances. See, for example, Kovačević, Yung and Pandher (1995). However, we were interested in seeing if the key results relating to the bias of the variance estimators would still hold if the CVs were lowered. Therefore, at the suggestion of a referee, we reran the simulation, increasing the number of PSUs drawn from 47 to 83,

since CVs of variance estimators are known to be approximately inversely related to the number of PSUs drawn. The PSUs were increased in such a way that the overall design was made self-weighting; this approach appeared to have the greatest effect on lowering the CVs. The second stage of sampling remained the same as before. Rerunning the simulation had the secondary benefit of roughly doubling the sample size, and thus, solving the convergence problems referred to in the last paragraph.

The results from the second run of the simulation are reported in Table 2. The last column in Table 2 shows the reduced number of discarded samples due to convergence problems. The fifth and sixth column of this table show that the CVs are significantly reduced to between 22.70% and 24.2% with the Jackknife consistently exhibiting slightly higher values. Now, as before, the percent relative bias in the point estimator is negligible, always being well under 1%. In the previous run, the percent relative biases for the Taylor estimator were always roughly -6%; here, they are always about -3%, again implying underestimation of the true variance. Once more, in the case of the GLS distance function, there is very little difference in the bias that results from using equation (3.1) versus (3.2). The percent relative bias in the Jackknife estimator (always roughly -1.5%) is consistently

Table 2  
Percent Relative Bias of the Point Estimators, and Percent Relative Bias and Percent CV of the Taylor and Jackknife Variance Estimators (Sample Size About 2000)

Distance Function		Percent Relative Bias Point Estimator	Percent Relative Bias Taylor Variance	Percent Relative Bias Jackknife Variance	Percent CV Taylor Variance	Percent CV Jackknife Variance	Number of Discarded Samples (From 4000)
GLS (Regression)		.02	-2.71 (eq 3.1) -2.61 (eq 3.2)	-1.43	23.03 (eq 3.1) 22.84 (eq 3.2)	23.29	0
Restricted GLS	( $L = 0$ , $U = 4$ )	.02	-2.61	-1.43	22.84	23.29	0
	( $L = .4$ , $U = 2$ )	.02	-2.61	-1.43	22.84	23.29	0
	( $L = .68$ , $U = 1.6$ )	.02	-2.61	-1.44	22.84	23.29	0
	( $L = .8$ , $U = 1.3$ )	.02	-2.75	-1.56	22.70	23.15	118
Raking Ratio		.25	-2.75	-1.15	22.84	23.43	0
Restricted RR	( $L = 0$ , $U = 4$ )	.17	-2.67	-1.36	22.84	23.30	0
	( $L = .4$ , $U = 2$ )	.16	-2.70	-1.42	22.84	23.29	0
	( $L = .68$ , $U = 1.6$ )	.31	-2.77	-0.49	22.83	24.20	0
	( $L = .8$ , $U = 1.3$ )	.27	-2.91	*	22.70	*	118
Modified Huang-Fuller	( $L = 0$ , $U = 4$ )	.02	-2.61	-1.43	22.84	23.29	0
	( $L = .4$ , $U = 2$ )	.02	-2.61	-1.43	22.84	23.29	0
	( $L = .68$ , $U = 1.6$ )	.02	-2.61	-1.44	22.84	23.29	0
	( $L = .8$ , $U = 1.3$ )	.02	-2.58	-1.36	22.73	23.18	116
Shrinkage-Minimization	( $L = 0$ , $U = 4$ )	.02	-2.61	-1.43	22.84	23.29	0
	( $L = .4$ , $U = 2$ )	.02	-2.61	-1.43	22.84	23.29	0
	( $L = .68$ , $U = 1.6$ )	.02	-2.61	-1.44	22.84	23.29	0
	( $L = .8$ , $U = 1.3$ )	.02	-2.61	-1.24	22.73	23.63	118



smaller in absolute value than that of Taylor. For the Jackknife estimator, there is one case (Restricted RR ( $L = .8$ ,  $U = 1.3$ )) where there were convergence problems; those results are omitted, indicated by a “\*”. Surprisingly, for both the Taylor and Jackknife, there is virtually no change in bias for the restricted distance functions as the bounds are made successively more tight. In fact, there seems to be very little difference in the percent relative bias across all of the distance functions, for both the Taylor and the Jackknife. Note that for the rerun of the simulation, the Monte Carlo errors ranged between .37% and 2.13%.

## 5. CONCLUSIONS

This paper focused on exploring the behaviour of point estimators and their corresponding Taylor and Jackknife variance estimators for a number of different distance functions available through calibration theory. Particular emphasis was given to those distance functions which allowed range restrictions to be imposed on the  $g$ -factors, eliminating the possibility of negative and high positive final weights. All of the point estimators which were investigated exhibited a negligible bias.

Both the Jackknife and Taylor variance estimators exhibited small underestimation of the true variance, although the Jackknife consistently had smaller biases (in absolute value) than the Taylor. The most striking result was that, for both Taylor and Jackknife, the biases remained roughly the same in the cases of extreme bounding on the  $g$ -factors as in the cases of less restrictive bounding. In general, however, caution should be exercised in the use of extreme bounds, due to the convergence problems that may be experienced, particularly when Jackknifing is used for variance estimation and the point estimators must be recalculated repeatedly. If the main objective of using the restricted distance functions is to eliminate the possibility of negative or high positive weights, then modest bounds on the  $g$ -factors should suffice.

As a final remark, it is interesting to note that roughly 97% of the computing time was spent Jackknifing while the remaining 3% was spent on Taylor linearization. This rather extreme difference in computation time may give the Taylor method an advantageous edge if measures of precision are required for a large number of domains. However, given recent developments in the computational efficiency of the Jackknife variance estimator (for example, the program WESVARPC (1995)), it may be possible to offset this imbalance. Even so, it should be noted that, at this time, WESVARPC has improved the computational efficiency for designs having only two PSUs per stratum, and poststratified estimators having only one dimension.

In conclusion, since our study does not conclusively show either variance estimator to be clearly superior and shows both to behave reasonably well for all distance functions, it is up to the user to decide which variance/ distance function combination best fits the system requirements.

## ACKNOWLEDGEMENTS

The authors would like to thank Chris Mohl for providing us with some of the computer code which we used in the simulation study. We would also like to thank an Associate Editor and two referees for useful comments on the earlier version of this paper.

## REFERENCES

- DEMING, W.E., and STEPHAN, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11, 427-444.
- DEVILLE, J.-C., and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- DEVILLE, J.-C., SÄRNDAL, C.-E., and SAUTORY, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.
- HIDIROGLOU, M.A., FULLER, W.A., and HICKMAN, R.D. (1980). SUPERCARP, Department of Statistics, Iowa State University, Ames, Iowa.
- HUANG, E.T., and FULLER, W.A. (1978). Nonnegative regression estimation for sample survey data. *Proceedings of the Social Statistics Section, American Statistical Association*, 300-305.
- KOVAČEVIĆ, M.S., YUNG, W., and PANDHER, G.S. (1995). Estimating the sampling variances of measures of income inequality and polarization – An empirical study. Methodology Branch Working Paper, HSMD-95-007E, Statistics Canada.
- SÄRNDAL, C.-E. (1982). Implications of survey design for generalized regression estimation of linear functions. *Journal of Statistical Planning and Inference*, 7, 155-170.
- SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- SINGH, A.C., and MOHL, C.A. (1996). Understanding calibration estimators in survey sampling. *Survey Methodology*, 22, 107-115.
- SINGH, M.P., DREW, J.D., GAMBINO, J.G., and MAYDA, F. (1990). *Methodology of the Canadian Labour Force Survey: 1984-1990*. Catalogue No. 71-526, Statistics Canada.
- STUKEL, D.M., and BOYER, R. (1992). Calibration estimation: An application to the Canadian Labour Force Survey. Methodology Branch Working Paper, SSMD-92-009E, Statistics Canada.
- WESVARPC (1995). Westat Inc., Rockville, Maryland.
- WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.
- YUNG, W., and RAO, J.N.K. (1996). Jackknife linearization variance estimators under stratified multi-stage sampling. *Survey Methodology*, 22, 23-31.





# An Application of Restricted Regression Estimation in a Household Survey

BODHINI R. JAYASURIYA and RICHARD VALLIANT<sup>1</sup>

## ABSTRACT

This paper empirically compares three estimation methods – regression, restricted regression, and principal person – used in a household survey of consumer expenditures. The three methods are applied to post-stratification which is important in many household surveys to adjust for under-coverage of the target population. Post-stratum population counts are typically available from an external census for numbers of persons but not for numbers of households. If household estimates are needed, a single weight must be assigned to each household while using the person counts for post-stratification. This is easily accomplished with regression estimators of totals or means by using person counts in each household's auxiliary data. Restricted regression estimation refines the weights by controlling extremes and can produce estimators with lower variance than Horvitz-Thompson estimators while still adhering to the population controls. The regression methods also allow controls to be used for both person-level and household-level counts and quantitative auxiliaries. With the principal person method, persons are classified into post-strata and person weights are ratio adjusted to achieve population control totals. This leads to each person in a household potentially having a different weight. The weight associated with the "principal person" is then selected as the household weight. We will compare estimated means from the three methods and their estimated standard errors for a number of expenditures from the Consumer Expenditure survey sponsored by the U.S. Bureau of Labor Statistics.

**KEY WORDS:** Calibration; Principal person method; Replication variance; Restricted regression.

## 1. INTRODUCTION

A signal problem in large household surveys is under-coverage of the target population often arising from differential response rates among population subgroups and frame deficiencies. Post-stratification is one method used at the estimation stage to reduce mean square errors based on information that affect the response variables. The estimator is constructed in such a way that the estimated total number of individuals falling into each post-stratum is equal to the true population count. Post-stratum population counts are typically available from an external census for numbers of persons but not always for numbers of households. If household estimates are needed, a single weight must be assigned to each household while using the person counts for post-stratification. Regression estimators of totals or means accomplish this by using person counts in each household's auxiliary data. Restricted regression estimation controls extreme weights and can produce estimators with lower variance than the Horvitz-Thompson estimator while still adhering to the population controls. An alternative used by some surveys is the Principal Person (PP) method (Alexander 1987) in which the household weight is based on the individual designated as the "principal person" in each household. Persons are classified into post-strata and person weights are ratio adjusted to achieve population control totals, leading to the possibility that each person in a household may have a different weight. The weight associated with the principal person is then assigned to the household. This *ad hoc* method is difficult to analyze theoretically. The regression estimators discussed in this

paper, while easily adjusting for the population under-count, automatically provide a household weight that is not based on any particular one of its members. Lemaître and Dufour (1987) address Statistics Canada's use of the regression estimator in this regard.

There are a growing number of precedents for the use of regression estimators in surveys both in the theoretical literature and in actual survey practice. Statistics Canada has incorporated the general regression estimator into its generalized estimation system (GES) software that is now used in many of its surveys (Estevao, Hidioglou and Särndal 1995). Fuller, Loughin and Baker (1993) discuss an application to the USDA Nationwide Food Consumption Survey. One of the attractions of regression estimation is that many of the standard techniques in surveys including the post-stratification estimator mentioned above are special cases of regression estimators. The regression estimator also more flexibly incorporates auxiliary data than other more common methods. In a household survey, for example, both person-level and household-level auxiliaries that can be qualitative or quantitative are easily accommodated. Other works related to regression estimation and post-stratification include Bethlehem and Keller (1987), Casady and Valliant (1993), Deville and Särndal (1992), Deville, Särndal and Sautory (1993) and Zieschang (1990).

In this study we compare the regression estimator with the PP estimator currently in use at the Bureau of Labor Statistics (BLS). Each estimator can be written in the form of a weighted sum of the sample values of the response variable. Then each weight is traditionally interpreted as the number of

<sup>1</sup> Bodhini R. Jayasuriya and Richard Valliant, U.S. Bureau of Labor Statistics, 2 Massachusetts Avenue, N.E., Room 4915, Washington, DC 20212, U.S.A.

individuals in the population who would have the corresponding value of the response variable. This interpretation requires that each weight be greater than or equal to one. The ordinary least-squares regression estimator has the disadvantage that it can produce non-positive weights. A number of ways are suggested in the literature on how to overcome this problem. Possibly the easiest is the method introduced by Deville and Särndal (1992) which can remove any negative weights as well as control extreme weights. The restricted regression estimators produced by these new weights are also compared to the original regression estimator and the PP estimator.

In Section 2, the three different estimators are presented. Section 3 is an application of these procedures to the Consumer Expenditure (CE) Survey at BLS – the same setting as in Zieschang (1990). We compare the coefficients of variation for a number of the survey target variables for the full population and for a number of domains. Section 4 provides a summary of our conclusions.

## 2. REGRESSION, CALIBRATION AND PRINCIPAL PERSON ESTIMATION

First, we give a brief introduction to the regression estimator. A sample  $s$  of size  $n$  is selected from a finite population  $U$  of size  $N$ . Let the probability of selection of the  $i$ -th unit be  $\pi_i$ . The sample could be two-stage and the unit could be either the primary sampling unit or the secondary sampling unit. There is no need here to complicate the notation with explicit subscripts for the different stages of sampling. Let the variable of interest be denoted by  $y$  and suppose that its value at the  $i$ -th unit,  $y_i$ , is observed for each  $i \in s$ . Assume the existence of  $K$  auxiliary variables  $x_1, x_2, \dots, x_K$  whose values at each  $i \in s$  are available. Define  $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{iK})'$ , for each  $i \in U$ , where  $x_{ik}$  denotes the value of the variable  $x_k$  at unit  $i$ . Let  $\mathbf{X} = (X_1, \dots, X_K)'$  denote the  $K$ -dimensional vector of known population totals of the variables  $x_1, x_2, \dots, x_K$ . The regression estimator is then motivated by the working model  $\xi$ :

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_K x_{iK} + \varepsilon_i \quad (2.1)$$

for  $i = 1, \dots, N$ . Here,  $\beta_1, \dots, \beta_K$  are unknown model parameters. The  $\varepsilon_i$  are random errors with  $E_\xi(\varepsilon_i) = 0$  and  $\text{var}_\xi(\varepsilon_i) = \sigma_i^2$  for  $i = 1, \dots, N$ . The term “working model” is used to emphasize the fact that the model is likely to be wrong to some degree. In the CE, the unit of analysis, indexed by  $i$ , is a consumer unit (CU), which is similar to a household and defined in more detail in Section 3. The value  $y_i$  might be the total food expenditures by the CU and the  $x_{ik}$ 's might be various CU characteristics like numbers of people of different ages, or CU income, that have an effect on the CU's expenditure on food. The variance of expenditures might be dependent on CU size so that having  $\sigma_i^2$  proportional to the number of persons in the CU might be reasonable. We include an intercept in some of our models by setting the first auxiliary variable,  $x_1$ , equal to 1.

A linear regression estimator of the population total of  $y$  is defined to be

$$\hat{y}_R = \hat{y}_\pi + (\mathbf{X} - \hat{\mathbf{x}}_\pi)' \hat{\boldsymbol{\beta}} \quad (2.2)$$

where  $\hat{y}_\pi$  denotes the  $\pi$ -estimator (or Horvitz-Thompson estimator) of the population total of  $y$ , i.e.,

$$\hat{y}_\pi = \sum_{i \in s} a_i y_i \quad (2.3)$$

with  $a_i = 1/\pi_i$ . Also,  $\hat{\mathbf{x}}_\pi = (\hat{x}_{1\pi}, \dots, \hat{x}_{K\pi})'$  is the vector of  $\pi$ -estimators of the population totals of the variables  $x_1, x_2, \dots, x_K$  and

$$\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \dots, \hat{\beta}_K)' = \left[ \sum_{i \in s} \frac{a_i \mathbf{x}_i \mathbf{x}_i'}{\sigma_i^2} \right]^{-1} \sum_{i \in s} \frac{a_i \mathbf{x}_i y_i}{\sigma_i^2}. \quad (2.4)$$

We assume that  $\sum_{i \in s} a_i \mathbf{x}_i \mathbf{x}_i' / \sigma_i^2$  is nonsingular. Even if model (2.1) fails to some degree,  $\hat{y}_R/N$  is a design consistent estimator of the population mean  $\bar{Y}$  irrespective of whether the assumed model is true or false. This is clear from (2.2). If  $\hat{y}_\pi/N$  and  $\hat{\mathbf{x}}_\pi/N$  are design consistent estimators of  $\bar{Y}$  and of  $\bar{\mathbf{X}}$ , the vector of population means of the auxiliaries, then the second term in  $\hat{y}_R/N$  converges to zero while the first converges to  $\bar{Y}$ . For more details, see Särndal, Swensson and Wretman (1992).

The regression estimator  $\hat{y}_R$  can also be expressed as a weighted sum of the sample  $y_i$ 's, which is a desirable feature for survey operations. It is easily seen that (2.2) can be re-written as  $\hat{y}_R = \sum_{i \in s} w_i y_i$  with

$$w_i = a_i \left[ 1 + (\mathbf{X} - \hat{\mathbf{x}}_\pi)' \mathbf{A}^{-1} \frac{\mathbf{x}_i}{\sigma_i^2} \right] \quad (2.5)$$

where  $\mathbf{A} = \sum_{i \in s} a_i \mathbf{x}_i \mathbf{x}_i' / \sigma_i^2$ . The weights do depend on the sample through the  $\mathbf{x}_i$ 's that are in the sample, but this is also true of many survey estimators, including the post-stratification estimator. However, these weights do not depend on the particular  $y$  variable being studied, implying that one set of  $w_i$  weights can be used for all estimates.

A mean per unit is estimated in the obvious way:  $\hat{\bar{y}}_R = \hat{y}_R / \hat{N}$  where  $\hat{N} = \sum_{i \in s} w_i$ . If we estimate the totals of the auxiliaries  $\mathbf{x}_i$ , then

$$\begin{aligned} \sum_{i \in s} w_i \mathbf{x}_i' &= \sum_{i \in s} \left[ a_i \mathbf{x}_i' + (\mathbf{X} - \hat{\mathbf{x}}_\pi)' \mathbf{A}^{-1} \frac{a_i \mathbf{x}_i \mathbf{x}_i'}{\sigma_i^2} \right] \\ &= \mathbf{X}', \end{aligned} \quad (2.6)$$

i.e., we reproduce the known population totals. This is also a characteristic of the post-stratification estimator.

The estimator of  $\boldsymbol{\beta}$  in (2.4) does not account for any correlation among the errors in model (2.1). In clustered



populations, units that are geographically near each other, *e.g.*, CU's in the same neighborhood, may be correlated. Using a full covariance matrix  $V$  may be more nearly optimal (*e.g.*, see Casady and Valliant 1993 and Rao 1994). Though use of a full covariance matrix  $V$  may lower the variance of  $\hat{\beta}$ , the elements of  $V$  will depend on the particular  $y$  being studied, and estimation of  $V$  is generally a nuisance. Consequently, it is interesting and practical to consider the simple case of  $V = \text{diag}(\sigma_i^2)$  that leads to (2.2). Note that when the design-variance  $\text{var}_p(\hat{y}_R)$  is estimated, it will be necessary to use a method that properly reflects clustering and other design complexities.

The regression estimator has the disadvantage that the weights can be unreasonably large, small or, even negative. The restricted calibration estimators of Deville and Särndal (1992), introduced next, add constraints to control the size of the weights. Calibration estimators are formed by minimizing a given distance,  $F$ , between some initial weight and the final weight, subject to constraints. The constraints can involve the available auxiliary variables thus incorporating them into the estimator. The regression estimator presented above is a special case of the calibration estimator in which  $F$  is defined to be the generalized least squares (GLS) distance function,

$$F(w_p, a_i) = \frac{a_i c_i}{2} \left( \frac{w_i}{a_i} - 1 \right)^2$$

for  $i = 1, \dots, n$ , with  $c_i$  a known, positive weight (*e.g.*,  $c_i = \sigma_i^2$  or  $c_i = 1$ ) associated with unit  $i$ , and  $w_i$ , the final weight. The total sample distance  $\sum_{i \in S} F(w_p, a_i)$  is minimized subject to the constraints,

$$\sum_{i \in S} w_i x_i = X. \quad (2.7)$$

In this form, the weights of the regression estimator of the population total of  $y$  given in (2.5) can be written as,

$$w_i = a_i g(c_i^{-1} \lambda' x_i) \quad (2.8)$$

for  $i = 1, \dots, n$  where

$$g(u) = 1 + u, \quad (2.9)$$

for  $u \in \mathcal{R}$  and  $\lambda$  is a Lagrange multiplier evaluated in the minimization process. The particular form of  $w_i$  with  $c_i = \sigma_i^2$  for the regression estimator was given in (2.5). To eliminate extremes, the weights can be refined by restricting  $g$  so that

$$g(u) = \begin{cases} L & \text{if } u < L - 1 \\ 1 + u & \text{if } L - 1 \leq u \leq U - 1 \\ U & \text{if } u > U - 1. \end{cases} \quad (2.10)$$

With this definition of  $g$ , the weights  $w_i$  satisfy

$$L < w_i/a_i < U \quad (2.11)$$

for  $i = 1, \dots, n$  so that  $L$  and  $U$  can be chosen in such a way as to reflect the desired deviation from the initial weights  $a_i$ . Choosing  $L > 0$  ensures that the weights are positive, and  $U$  is picked to be appropriately small to prohibit large weights. The restricted regression weights must be solved for iteratively; one easily programmed algorithm is given in Stukel and Boyer (1992). Another method of restricting weights is ridge regression as used by Bardsley and Chambers (1984).

In most household surveys, post-stratification serves primarily as an adjustment for under-coverage of the target population by the frame and the sample. In the U.S., there are few reliable population counts of households to use in post-stratification. Consequently, population counts of persons are usually used for the post-strata control totals. This disagreement in the unit of analysis (the household) and the unit of post-stratification (the person) when a household characteristic is of interest led to the development of the PP method that is used in the CE and Current Population Surveys.

In the PP method described in Alexander (1987), a household begins the weighting process with a single base weight,  $a_i$ , that is then adjusted for non-response. The adjusted weight is assigned to each person in the household and the person weights are then further adjusted to force them to sum to known population controls of persons by age, race, and sex. This last adjustment can result in persons having different weights within the same household. The household is then assigned the weight of the person designated as the "principal person" in the household. This method has an element of arbitrariness and is difficult to analyze mathematically. The intent of this research was not to see if the PP method could be improved upon, but rather to use the current implementation of PP as a convenient baseline for measuring the performance of other estimators.

The regression and restricted regression estimators can be formulated in such a way that population person controls are satisfied, all persons in a household retain the same weight, and no arbitrary choice among person weights is needed to assign a household weight. This is accomplished by defining the auxiliary variables at the household level. For example, if there were three age post-strata and household  $i$  has 1, 0, and 2 persons in these post-strata, the auxiliary data vector would be  $x_i = (1, 0, 2)'$ . Note that this formulation is different from Lemaître and Dufour (1987) who defined the auxiliary variables at the person level and assigned the average of the household data – (1/3, 0, 2/3) in the example – to each person. Those authors used this "average" method because they were interested in estimates both for persons, *e.g.*, number employed, and for households, *e.g.*, economic families. We, on the other hand, need only a household weight since our target variables (*i.e.*,  $y$ ) like shelter or utility expenditures are collected at the household level.

### 3. AN APPLICATION

We compare the three estimators (*i.e.*, regression, restricted regression (with  $L = .5$ ,  $U = 4$ ), and principal person) by an

application to the estimated means and their estimated standard errors for a number of expenditures from the CE Survey sponsored by the Bureau of Labor Statistics.

The CE Survey gathers information on the spending patterns and living costs of the American consumers. There are two parts to the survey, a quarterly interview and a weekly diary survey. The Interview Survey collects detailed data on the types of expenditures which respondents can be expected to recall for a period of three months or longer (*e.g.*, property, automobiles, major appliances) – an estimated sixty to seventy percent of total household expenditures. The Diary Survey is completed at home by the respondent family for two consecutive 1-week periods and collects data on all the expenses of the family in that time period. The sample is selected in two stages with geographic primary sampling units at the first stage and households at the second.

We evaluated the estimators described above for a number of expenditures from the Interview Survey. Data collected during the second quarter of 1992 consisting of  $n = 5156$  CU's were used. The CE Survey's primary unit of analysis is the consumer unit, an economic family within a household. A consumer unit (CU) consists of individuals in the household who share expenditures. Thus, there may be more than one CU in a household.

Five different sets of auxiliary variables ( $x_i$ 's in the notation of Section 2) were studied. They were chosen by testing the adequacy of model (2.1) for the selected expenditures with different combinations of the available auxiliary variables. Combinations of auxiliaries were identified in which each estimated regression coefficient was significant in an ordinary least squares regression at the 5% level. A key step that substantially improved the fit of the models was simply including an intercept. Factored into the selection of auxiliaries was also the knowledge that the survey has more under-coverage of Blacks than non-Blacks and that this needed to be accounted for by post-stratification. We viewed this method of variable selection as exploratory and, consequently, a number of combinations were studied to determine which set produced the best estimators of mean expenditures. The 56 post-strata based on age/race/sex currently in use in the CE were included. (The 56 are routinely collapsed in actual CE operations because of small sample sizes in some cells.) Other variables that were statistically significant in various combinations were region (NE, MW, S, W), urbanicity (urban/rural) by region, age of reference person of the CU (< 25, 25-34, 35-44, 45-64, 65+), household tenure (owner/renter), income before taxes of the CU, and the 56 post-strata collapsed by sex and some of the age categories to form 10 age/race categories. Based on this information, weights (2.8) were computed using  $g$  given in (2.9) – regwts – and (2.10) – calwts. For both the regression and restricted regression weights, we set  $a_i$  equal to the adjusted base weight, *i.e.*,  $1/\pi_i$  times a non-response adjustment. In order for the matrix  $A$  in Section 2 to be nonsingular, one of the categories in some auxiliaries, like region, was omitted from each  $x_i$ . For this application, the population totals necessary

to evaluate  $X = (X_1, \dots, X_K)'$  were obtained mostly from the Statistical Abstract of the United States (1993) whose sources are the 1990 Census figures and the Current Population Reports published by the U.S. Bureau of the Census. When an intercept is used, the appropriate control total for that variable is the number of CU's in the population for which we used the PP estimate as a surrogate. The combinations of auxiliaries used to form the different weights are given in Table 1. Regwts0, with 56 age/race/sex post-strata uses the largest number of post-strata. The 56 are the starting point for the PP method but are usually collapsed to 30-40 because of small cell sizes. When computing calwts0, those 56 post-strata were collapsed to 45 since the constraints imposed by the  $L$  and  $U$  bounds could cause singularity in the matrix based algorithm.

**Table 1**  
Weights and Their Corresponding Auxiliary Variables

Weights	Auxiliary Variables	$K$
regwts0	Age/race/sex	56
regwts1	Intercept, age/race/sex, region, urban $\times$ region	18
regwts2	Intercept, age/race/sex, region, urban $\times$ region, age of reference person, housing tenure, family income before taxes	24
calwts0	Age/race/sex	45
calwts1	Intercept, age/race/sex, region, urban $\times$ region	18
calwts2	Intercept, age/race/sex, region, urban $\times$ region, age of reference person, housing tenure, family income before taxes	24
calwts3	Intercept, age/race/sex, region, urban $\times$ region, family income before taxes (truncated at \$500,000)	19
calwts4	Intercept, age/race/sex, region, urban $\times$ region, age of reference person, housing tenure	23
PP	Age/race/sex	56 <sup>1</sup>

<sup>1</sup> The initial set of 56 is usually collapsed to 30-40 because of small sample sizes in some cells.

### 3.1 Comparisons of Weights

A variety of comparisons of weights produced by the different methods were made, only a few of which can be mentioned here. Figure 1 shows plots of the PP weights, regwts0, calwts0, and calwts1 versus the adjusted base weights. For PP and regwts0, the adjustments to go from  $a_i$  to  $w_i$  are much more variable than for calwts0 and calwts1, which employ the  $L = 0.5$  and  $U = 4$  restrictions. High variability among the  $w_i$  can lead to expenditure estimates with high variance and to poor confidence interval coverage since large sample normality may not hold. Even though (2.11) implies that  $a_i/2 < w_i < 4a_i$  for each  $i$  for the calwts, the lower right panel in Figure 1 shows that the calwts1 satisfy  $a_i/2 < w_i \leq 2a_i$  for each  $i$ . Thus, setting  $U = 2$  or 3 would have little effect on calwts1. Calwts0 would have been slightly affected by setting  $U = 2$  since a few points were outside the upper reference line. The upper two panels indicate that the PP weights and regwts0 do not conform to the restriction  $a_i/2 < w_i < 2a_i$ .



(Reference lines correspond to  $L = .5$  and  $U = 2$ )

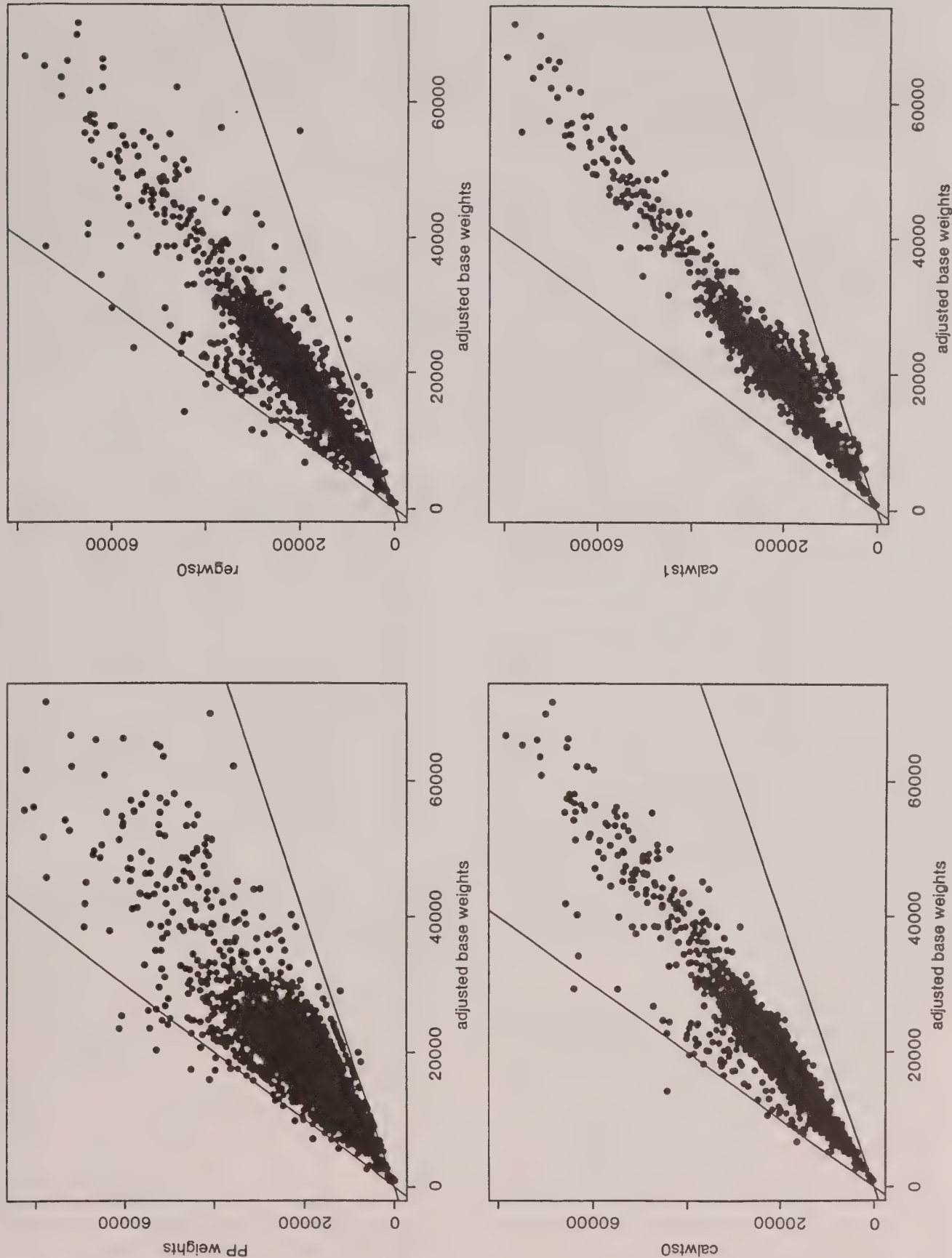


Figure 1. Four sets of weights plotted versus adjusted base weights

The concern about negative regression weights was minor in the application. In the full sample, only one CU had a negative weight for regwts1 and regwts2 while regwts0 had no negative weights. However, in the replicates used for variance estimation, described in Section 3.2, 2 or 3 CU's did have negative weights in many replicates so that using the  $L$  restriction was more important there.

### 3.2 Precision of Estimates from the Different Methods

Although comparison of weights is instructive, the methods must ultimately be judged based on the level of estimated CU means and their precision. The standard errors of these estimators were computed via the method of balanced half sampling (BHS) using 44 replicates as currently implemented in the CE for the PP estimator. The BHS estimator is constructed to reflect the stratification and the clustering that is used in the CE. A half sample is constructed in a prescribed way (McCarthy 1969) to contain one half of the first-stage sample units in a survey. Defining the mean per CU based on CU's in half-sample  $\alpha$  to be  $\hat{y}_{R(\alpha)}$  and that for the full sample to be  $\hat{y}_R$ , the BHS estimate of variance is  $V_{BHS}(\hat{y}_{R(\alpha)}) = \sum_{\alpha=1}^{44} (\hat{y}_{R(\alpha)} - \hat{y}_R)^2 / 44$ . To compute each  $\hat{y}_{R(\alpha)}$ , the same estimation steps used for the full sample are repeated for the CU's in the half-sample. As the expenditure estimates from the CE Survey are published for various inter domains of interest, we computed the means and the standard errors for a few chosen domains as well. For each of these, the coefficient of variation (cv) was computed and then its ratio to the cv of the PP weight estimate was calculated.

For each type of weight, if the ratio of each expenditure cv to that of the PP weights is less than one, an improvement over the PP estimate is indicated since, for all the weights, the expenditure mean estimates were very close to those of the PP estimates. We computed the ratios of cv's and the ratios of means for each of the sets of weights described in Table 1, for each of the chosen expenditures, and for each of the following domains:

- (1) Age of Reference Person: < 25, 25-34, 35-44, 45-54, 55-64, 65+
- (2) Region: NE, MW, S, W
- (3) Size of CU: 1, 2, 3, 4, 5+
- (4) Composition of Household: Husband and wife only, Husband and wife + children, Other Husband and wife, One parent + at least one child < 18, Single person and other CU's
- (5) Household Tenure: Owner, Renter
- (6) Race of Reference Person: Black, Non-Black.

We will discuss only domains (1) – (3) here. In addition, ratios for all CU's, *i.e.*, the total across the domains, were computed for each expenditure and are shown in Table 2. For All Expenditures, regwts2, calwts2, and calwts3, with ratios of .79, .78, and .75, provide substantial reduction in cv compared to PP. For less aggregated expenditures regwts1 or calwts1 provide reasonably consistent improvements over PP

**Table 2**

Ratios to PP cv of cv's for the Different Weighting Methods  
The Minimum Ratio is Highlighted in Each Row

Expenditure	regwts			calwts				
	0	1	2	0	1	2	3	4
All expenditures	0.98	0.90	0.79	0.98	0.90	0.78	0.75	0.87
Shelter	0.93	0.85	0.75	0.93	0.85	0.74	0.72	0.84
Utilities	1.08	1.03	0.94	1.07	1.03	0.88	0.91	0.92
Furniture	1.08	1.21	3.52	1.06	1.21	2.58	2.57	1.17
Major appliances	1.08	1.06	1.04	1.06	1.08	1.09	1.00	1.03
All vehicles	0.90	0.89	0.98	0.91	0.89	0.98	0.97	0.90
New cars, trucks	0.95	0.91	1.01	0.96	0.91	1.02	1.02	0.91
Used cars, trucks	0.98	0.94	0.96	0.97	0.94	0.97	0.96	0.95
Gasoline, motor oil	1.17	1.11	1.03	1.12	1.10	0.99	0.94	1.10
Health care	1.05	0.97	0.86	1.07	0.97	0.85	0.87	0.94
Education	0.92	0.93	1.04	0.91	0.93	1.06	1.07	0.88
Cash contributions	1.01	1.02	1.28	1.01	1.02	1.30	1.29	1.03
Personal insurance, pensions	1.00	0.97	1.64	1.01	0.98	1.24	0.98	0.95
Life, other personal insurance	1.08	1.02	1.53	1.08	0.98	1.38	1.33	1.01
Pensions, social security	1.00	0.99	1.75	1.01	0.99	1.34	1.06	0.97

without the losses incurred by some of the other weights for expenditures like Furniture, Personal insurance and pensions, and its sub-category Pensions and social security.

Trellis plots (Cleveland 1993) of the cv and mean ratios for calwts0 and calwts1 are given in Figures 2-4. Calwts0 is pictured because it is the nearest calibration equivalent to the current method of post-stratification. Calwts1 appears to be the best of the alternatives we have examined in the sense of improving the All Expenditures estimates while providing consistent performance for individual expenditure groups. In each panel of the plots a vertical reference line is drawn at 1, the point of equality between the calibration results and those for the PP method. The lower row in each plot presents ratios of means from calwts0 and calwts1 to the PP means and illustrates that with a few exceptions the levels of the means from the two restricted regression choices are about the same as from PP.

The two calibration choices, in the main, improve cv's compared to PP, *i.e.*, cv ratios tend to be less than 1, for most domains and expenditures, and calwts1 is somewhat better than calwts0. For the age-of-reference-person domains < 25 and 65+, for example, 12 of the 15 expenditures have calwts1 ratios of less than 1. For CU sizes 1-4 the numbers of cv ratios less than or equal to 1 are 12, 9, 9, and 11. There are exceptions, of course. For the South region only 6 of 15 expenditures have calwts1 cv ratios less than or equal to 1.

Calwts2 and calwts3, which used family income before taxes as one of the auxiliaries, had somewhat erratic performance for domains, sometimes making major improvements over PP but occasionally showing serious losses. This is connected to the nature of the family income variable itself. For the entire data set of 5156 CU's, income before taxes was positive for 4698 CU's, zero for 450 CU's and negative for 8 CU's. The zeroes are incomplete income reporters while the negatives are for families that had business losses added to other income. In either case, these CU's vitiate the usefulness



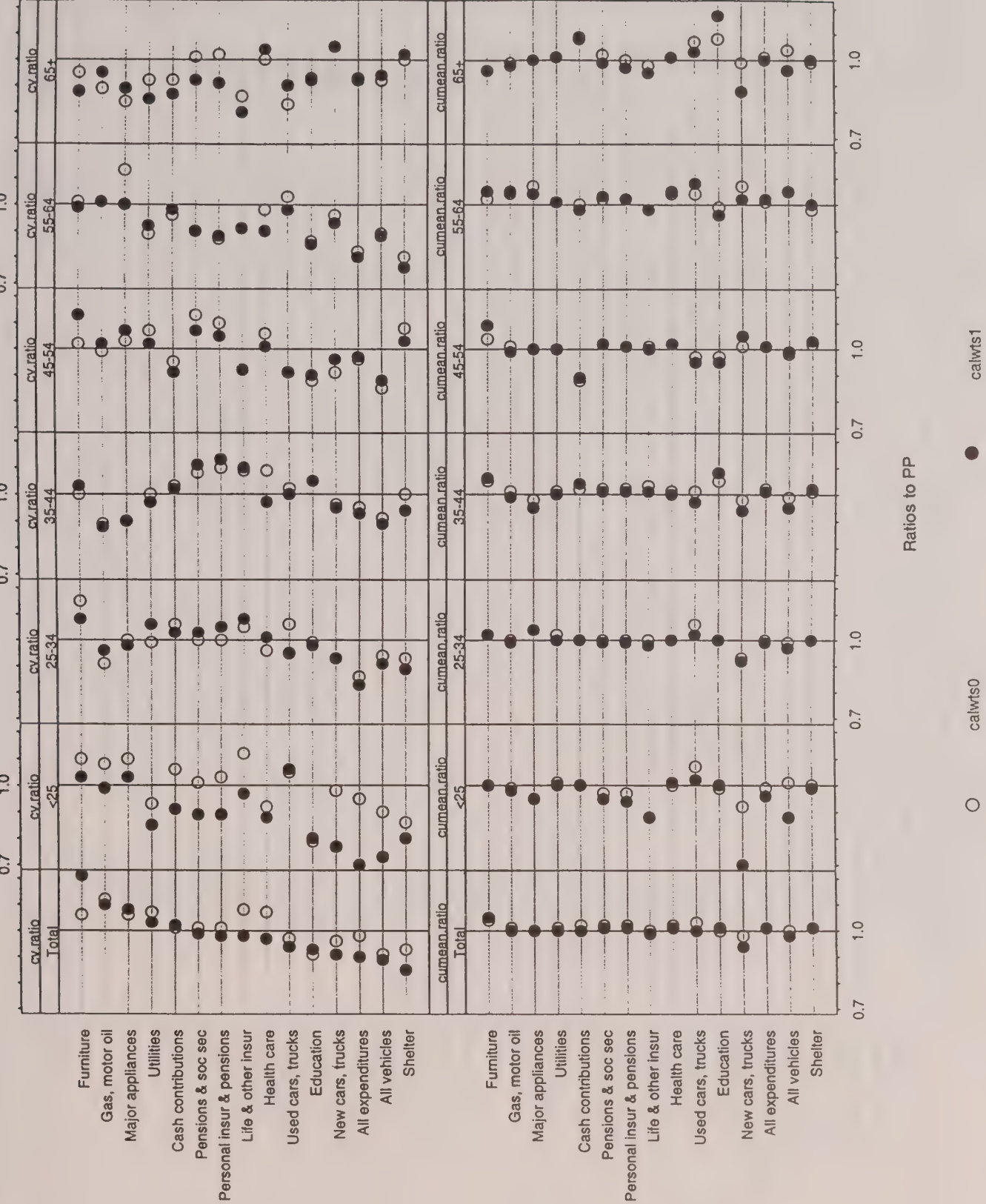


Figure 2. Ratios to PP of cv's and means for two weighting methods by age of reference person

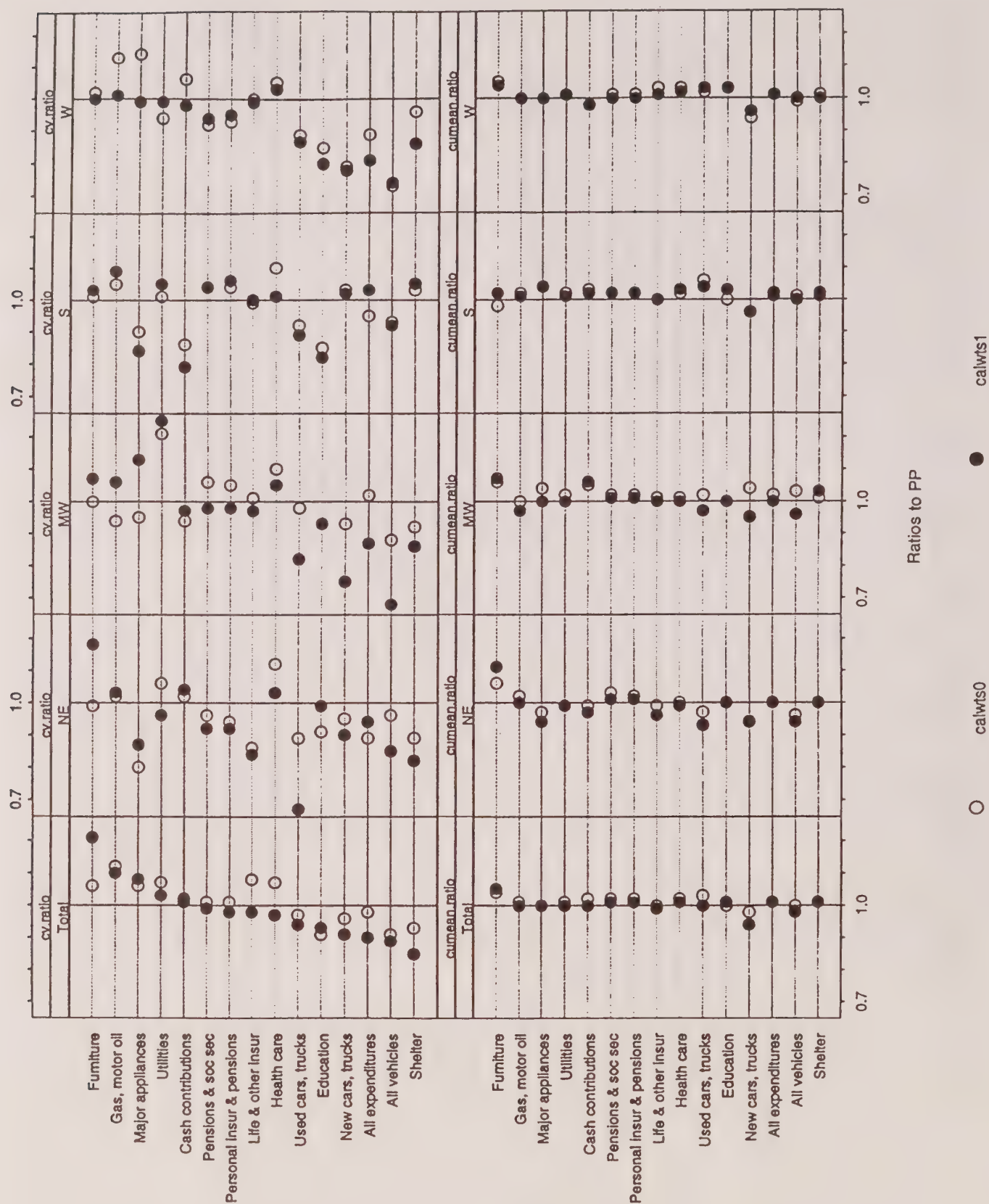


Figure 3. Ratios to PP of cv's and means for two weighting methods by region



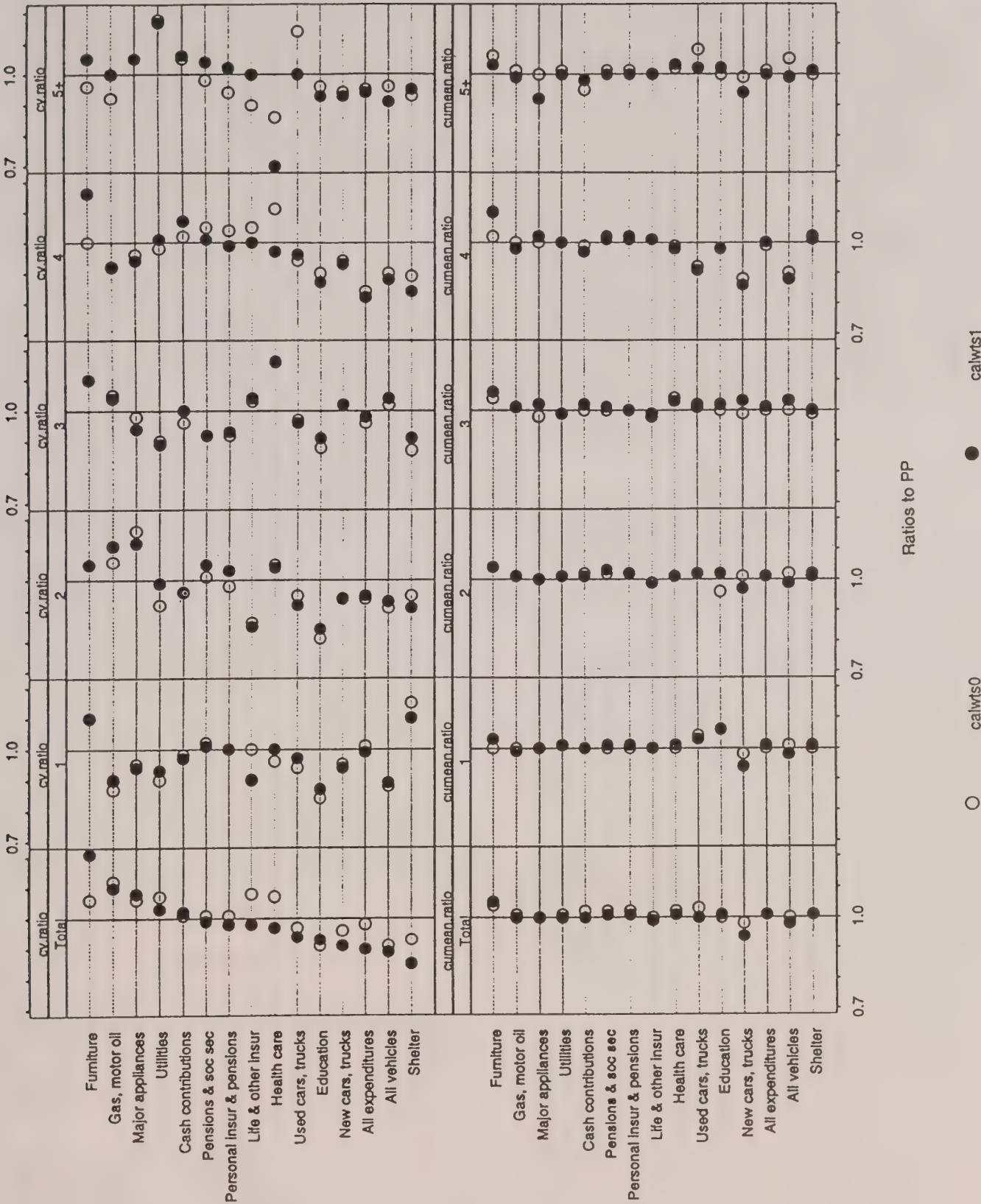


Figure 4. Ratios to PP of cv's and means for two weighting methods by size of CU

of this variable in predicting expenditures. Perhaps, use of another measure of income combined with item imputations for missing incomes would improve calwts2 and calwts3 for domain estimation.

Taking all of the above into consideration, regwts1, calwts1 and calwts4 are efficient choices in this application. Calwts1 has the advantage of non-negative weights over regwts1. Since calwts4 requires 23 auxiliary variables as opposed to calwts1's 18, calwts1 is the more parsimonious choice. Subsequent to the analysis discussed here, we performed a similar study using a full year's data for both the Interview and Diary Surveys for 1990. Results were similar to those reported here and a final set of 24 auxiliaries was adopted based on number of persons by age, race, sex, region, urban  $\times$  region, and number of CU's by tenure, and an intercept. The conversion of CE estimation to restricted regression is now underway.

#### 4. CONCLUSION

The objective of this study was to investigate methods for deriving household weights that did not depend on the weight of one single member of the household. Different types of weights based on the regression estimation procedure were presented and their relative merits evaluated. Regression estimation incorporates the current survey post-stratification methods in which the weighted sum of the persons in each post-stratum is forced to be equal to an independent census count of that number. This is accomplished via auxiliary variables that are incorporated into the regression model. It also automatically produces for each sample household a weight that does not depend on any single one of its members.

We studied eight types of weights that came from five different regression models. In order to eliminate the undesirable negative weights that can result from ordinary least-squares regression estimation, restricted regression estimators were adapted to the present problem. Restricted regression has the flexibility to restrict the possible deviation of each final weight from its base weight while adhering to the properties discussed above. This, in particular, allows the constraint of positive weights. The restricted regression weights are easily computed via matrix-oriented software like S-Plus™ or SAS/IML™.

Restricted regression, and more generally, restricted calibration have a number of attractive features for household surveys, like the one studied here, but also for surveys of other types of units like hospitals, schools, or business establishments where a variety of auxiliary data may be available. Given past data on target variables, standard model building procedures can be used for the selection of auxiliary variables. The properties of regression estimation can be used to choose the predictors optimally in order to reduce the redundancy of information that gets incorporated into the survey estimation procedure. This is one of the greatest advantages of using an estimator that has a vast and tested literature behind it. Good

predictors may include qualitative variables, *e.g.*, age, race, type of hospital (general medical, psychiatric, *etc.*), type of business (manufacturing, retail trade, *etc.*) that might be often used in stratification or post-stratification. The predictors can also be quantitative variables like family income, annual sales, number of students at different levels, or the number of inpatient days to name but a few. In our application, including an intercept also led to noticeably smaller standard errors of survey estimates. The regression approach also allows data at different levels to be easily incorporated in estimation. In the household survey studied here, auxiliaries on both persons and households were included.

The immense flexibility of regression gives practitioners options they might not otherwise have. If new, pertinent predictor variables become available, software for regression estimation can accommodate them simply by changing the matrix of auxiliaries and vector of population controls. Software that is rigidly written to perform only post-stratification or ratio estimation with a single auxiliary, for example, might have to undergo a major overhaul to change the estimator. Of course, if the estimator is one of the less general post-stratification or the ratio types, regression software will often handle it as a special case. In the United States, an extremely large continuing household survey is being contemplated (Love, Alexander and Dalzell 1995) that will provide very precise estimates of many characteristics that may be used as control totals in smaller surveys. The restricted regression approach positions the CE Survey to smoothly incorporate such new data in estimation should it become available.

#### ACKNOWLEDGMENTS

Any opinions expressed are those of the authors and do not represent policy of the Bureau of Labor Statistics. We are grateful to the referees and editor for their useful comments.

#### REFERENCES

- ALEXANDER, C.H. (1987). A class of methods for using person controls in household weighting. *Survey Methodology*, 13, 183-198.
- BARDSLEY, P., and CHAMBERS, R.L. (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics*, 33, 290-299.
- BETHLEHEM, J.G., and KELLER, W.J. (1987). Linear weighting of sample survey data. *Journal of Official Statistics*, 3, 141-153.
- CASADY, R.J., and VALLIANT, R. (1993). Conditional properties of post-stratified estimators under normal theory. *Survey Methodology*, 19, 183-192.
- CLEVELAND, W.S. (1993). *Visualizing Data*. Summit, New Jersey: Hobart Press.



- DEVILLE, J.-C., and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- DEVILLE, J.-C., SÄRNDAL, C.-E., and SAUTORY, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.
- ESTEVAO, V., HIDIROGLOU, M.A., and SÄRNDAL, C.-E. (1995). Methodological principles for a generalized estimating system at Statistics Canada. *Journal of Official Statistics*, 11, 181-204.
- FULLER, W.A., LOUGHIN, M.M., and BAKER, H.D. (1993). Regression weighting for the 1987-1988, Nationwide Food Consumption Survey. Unpublished report submitted to the United States Department of Agriculture.
- LEMAÎTRE, G., and DUFOUR, J. (1987). An integrated method for weighting persons and families. *Survey Methodology*, 13, 199-207.
- LOVE, S., ALEXANDER, C.H., and DALZELL, D. (1995). Constructing a major survey – operational plans and issues of continuous measurement. *Proceedings of the Section on Survey Research Methods, American Statistical Association*. To appear.
- McCARTHY, P.J. (1969). Pseudo-replication: half-samples. *Review of the International Statistical Institute*, 37, 239-264.
- RAO, J.N.K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics*, 10, 153-165.
- SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- STUKEL, D.M., and BOYER, R. (1992). Calibration estimation: an application to the Canadian Labor Force Survey. Working Paper, SSMD-92-009E, Ottawa: Statistics Canada.
- ZIESCHANG, K.D. (1990). Sample weighting methods and estimation methods in the Consumer Expenditure Survey. *Journal of the American Statistical Association*, 85, 986-1001.





# A Transformation Method for Finite Population Sampling Calibrated With Empirical Likelihood

GEMAI CHEN and JIAHUA CHEN<sup>1</sup>

## ABSTRACT

In this paper, we study a confidence interval estimation method for a finite population average when some auxiliary information is available. As demonstrated by Royall and Cumberland in a series of empirical studies, naive use of existing methods to construct confidence intervals for population averages may result in very poor conditional coverage probabilities, conditional on the sample mean of the covariate. When this happens, we propose to transform the data to improve the precision of the normal approximation. The transformed data are then used to make inference on the original population average, and the auxiliary information is incorporated into the inference directly, or by calibration with empirical likelihood. Our approach is design-based. We apply our approach to six real populations and find that when transformation is needed, our approach performs well compared to the usual regression method.

KEY WORDS: Finite population; Sampling; Confidence interval; Transformation; Empirical likelihood.

## 1. INTRODUCTION

Let  $(x_i, y_i)$ ,  $i = 1, 2, \dots, N$  be values associated with  $N$  units in a finite population. For unit  $i$ ,  $y_i$  is the variable of interest and  $x_i$  is an auxiliary variable. One of the most extensively studied finite population problems is the estimation of the population average  $\bar{y} = (y_1 + \dots + y_N)/N$  (or total  $N\bar{y}$ ) under various sampling schemes. We shall focus on the simple random sampling scheme in this paper, because the nature of the problems we want to study can be better seen from this scheme and the results obtained here can be easily generalized into other sampling schemes of which the simple random sampling scheme is the building block.

It is often true that some information about the auxiliary variable  $x$  is known and can be used to make inference about  $\bar{y}$ . For example, let  $S = \{1, \dots, i, \dots, N\}$  and let  $s \subset S$  be a simple random sample of size  $n$ . When  $\bar{x} = (x_1 + \dots + x_n)/n$  is known, and  $x$  and  $y$  are correlated, the population average  $\bar{y}$  can be estimated by the ratio estimator  $\hat{\bar{y}} = (\bar{y}_s/\bar{x}_s)\bar{x}$ , or by the regression estimator  $\hat{\bar{y}} = \bar{y}_s + b(\bar{x} - \bar{x}_s)$ , where  $\bar{x}_s$  and  $\bar{y}_s$  are the sample averages of  $x$  and  $y$ , respectively, and  $b = \sum (x_i - \bar{x}_s)(y_i - \bar{y}_s) / \sum (x_i - \bar{x}_s)^2$ .

Under very general conditions, both the ratio estimator and the regression estimator are asymptotically normal; see Scott and Wu (1981), Bickel and Freedman (1984), and Theorem 2.1 of Section 2. Hence, if  $v$  is a carefully chosen estimator of the variance of  $\hat{\bar{y}}$ , the standardized variable  $(\hat{\bar{y}} - \bar{y})/\sqrt{v}$  is customarily treated to have the standard normal distribution. Therefore, if  $z_\alpha$  denotes the upper  $\alpha$ -percentile of the standard normal distribution, then

$$(\hat{\bar{y}} - z_\alpha \sqrt{v}, \hat{\bar{y}} + z_\alpha \sqrt{v}) \quad (1.1)$$

will produce an approximate  $100(1 - 2\alpha)\%$  confidence interval for  $\bar{y}$ .

Confidence interval (1.1) is widely used in practice. However, problems arise when it is applied to certain populations. Royall and Cumberland (1981a, 1981b, 1985) studied the ratio and regression estimators and applied them to six real populations where strong correlations between  $x$  and  $y$  seemed to exist. (See Section 3 for a summary of the six populations.) Various estimators of the variance of  $\hat{\bar{y}}$  were used. It was found that the actual conditional coverage rate of the confidence interval (1.1), conditional on  $\bar{x}_s$ , depended heavily on the size of  $\bar{x}_s$  and were usually much lower than the claimed coverage rate, even with the most adaptive variance estimator. For example, the 95% confidence interval for a population named Counties 70 had a conditional coverage rate 76% with the jackknife variance estimator when  $\bar{x}_s$  was small, and the conditional coverage rate could go as low as 50% with other variance estimators.

The above mentioned studies point to the need to construct confidence intervals that "will live up to their name" (Royall and Cumberland 1985, p. 359). However, up to now there has been little progress made in this direction. In this paper, we present some results from studying an alternative procedure for constructing confidence intervals and from applying it to the six populations studied by Royall and Cumberland and many others. As will be shown in Section 3, the conditional coverage rate of our confidence intervals is more accurate.

Two important ideas, namely, transformation and empirical likelihood, are used simultaneously to attack the problems encountered by Royall and Cumberland in particular, and to develop a new procedure in general. As explained in Cochran (1977, p. 150), the preference in sample survey theory is to make, at most, limited assumptions about the frequency

<sup>1</sup> Gemai Chen, Department of Mathematics and Statistics, University of Regina, Regina, Saskatchewan, Canada, S4S 0A2; Jiahua Chen, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada, N2L 3G1.

distribution followed by the data in the sample. However, ratio or regression estimator can help obtain increased precision by taking advantage of the correlation between  $y_i$  and  $x_i$ . This, of course, can be described by some assumption(s), such as an approximate linear relationship between  $y$  and  $x$ . Although almost no further assumptions are necessary to use the ratio or regression approach, the procedure (1.1) is clearly based on a normal approximation. But as it is well known, the normal approximation can be very poor when the population distribution is severely skewed and the sample size is small. In terms of procedure (1.1), the closer the estimator distribution is to the normal, the better one can construct confidence intervals. If the population distribution is severely skewed, a transformation may produce a population distribution that is at least more symmetric, so that the normal approximation for the estimator is more accurate.

When using the ratio and regression estimators, knowing  $\bar{x}$  is crucial to gain improvement over the use of sample mean. In our proposed procedure, the complete information about the auxiliary variable  $x$  can be incorporated. But if  $\bar{x}$  is the only auxiliary information available, it is difficult to use this information directly when a transformation is involved, because any non-linear transformation obscures the link between  $\bar{x}$  and  $\bar{y}$ . In this second case, we find the method of empirical likelihood very helpful in solving our problem; see particularly Owen (1988, 1990) and Chen and Qin (1992) for references. The empirical likelihood method in this situation can also be regarded as a calibration method as discussed in Deville and Särndal (1992). This approach rescues us from losing information about  $x$  after transforming the data.

There have been many discussions on how to use transformations to make better inference on the transformed scale (Box and Cox 1964; Carroll and Ruppert 1988; Calvin and Sedransk 1991, and the references therein). There have also been some studies on how to make inference on the original scale, after a transformation is applied (Carroll and Ruppert 1984; Elliott 1977). What is new with our procedure is the attempt to link the above two steps by combining transformation with auxiliary information and/or by applying empirical likelihood method when necessary.

The details of our procedure are given in Section 2. Then our procedure is applied to the six populations studied by Royall and Cumberland in Section 3. The validity of our procedure in an arbitrary setting is demonstrated in Section 4 and some comments are made at the end of the paper.

## 2. THE NEW PROCEDURE

As mentioned in the last section, a problem with the confidence interval (1.1) is that it will fail if the distribution of  $(\hat{y} - \bar{y})/\sqrt{v}$  is severely asymmetric and far from the normal distribution. The problem can be inherited from the skewness of the population distribution. When the skewness is severe, a central confidence interval procedure like (1.1) is doomed to fail. The basic model employed by Royall and Cumberland (1981a, 1981b, 1985) is

$$y_i = \alpha + \beta x_i + \epsilon_i, \quad (2.1)$$

with  $E(\epsilon_i) = 0$ ,  $V(\epsilon_i) = \sigma^2$  and  $\text{Cov}(\epsilon_i, \epsilon_j) = 0$ , for  $i \neq j$ . It is easy to find that for the six real populations studied by Royall and Cumberland, the corresponding error distributions are very skewed. These observations lead us to consider transforming the variables  $y$  and/or  $x$ , and consider the model

$$h(y_i) = \alpha + \beta g(x_i) + \sigma \epsilon_i, \quad (2.2)$$

where  $h(\cdot)$  and  $g(\cdot)$  are two monotone functions. There are many families of transformations suggested in the literature. One commonly used family is the Box-Cox power transformation family defined by

$$f(x, \lambda) = \begin{cases} (x^\lambda - 1)/\lambda & \text{when } \lambda \neq 0, \\ \log(x) & \text{when } \lambda = 0. \end{cases}$$

Model (2.1) is a special case of (2.2) when both  $h$  and  $g$  equal  $f(x, 1)$ .

The choice of transformations in model (2.2) might be suggested by an examination of the sample  $x$ 's and  $y$ 's based on a possible model relationship, or by our subject knowledge about the population under investigation. For example, for the six populations discussed in Royall and Cumberland, the population distributions are severely skewed towards the right which can be learned from the nature of the finite populations. Therefore, a log transformation may make them all less skewed. Other more objective methods of choosing transformations are discussed in Section 4.

We emphasize that models (2.1) and (2.2) are used here to motivate transformations, point estimators, or confidence interval procedures. Our study of conditional coverage rates will, however, be based on the probability measure generated by the design, as in Royall and Cumberland (1985). For this purpose, we embed our finite population in a sequence of populations indexed by  $k$ . This means that a sub-index  $k$  is needed to write  $N = N_k$  and  $n = n_k$ , etc., but for simplicity, we will suppress the index  $k$  if there is no possibility for confusion.

Let  $v_i = h(y_i)$ ,  $u_i = g(x_i)$ ,  $\bar{v}_N = N^{-1} \sum_{i=1}^N v_i$  and  $\bar{u}_N = N^{-1} \sum_{i=1}^N u_i$ . Define

$$\beta_N = \frac{\sum_{i=1}^N (u_i - \bar{u}_N) v_i}{\sum_{i=1}^N (u_i - \bar{u}_N)^2},$$

$$\alpha_N = \bar{v}_N - \beta_N \bar{u}_N,$$

$$e_i = v_i - (\alpha_N + \beta_N u_i),$$

$$\sigma_N^2 = \frac{1}{N-1} \sum_{i=1}^N e_i^2.$$



Suppose  $s \subset S$  is a simple random sample of size  $n$ . We similarly define

$$\hat{\beta} = \frac{\sum_{i \in s} (u_i - \bar{u}_s) v_i}{\sum_{i \in s} (u_i - \bar{u}_s)^2},$$

$$\hat{\alpha} = \bar{v}_s - \hat{\beta} \bar{u}_s,$$

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i \in s} (v_i - \hat{\alpha} - \hat{\beta} u_i)^2,$$

where  $\bar{u}_s$  and  $\bar{v}_s$  are the sample averages.

Denote the inverse function of  $h(\cdot)$  by  $h^{-1}(\cdot)$ . Then the fitted value of  $y_i$  is

$$\hat{y}_i = h^{-1}(\hat{\alpha} + \hat{\beta} u_i). \quad (2.3)$$

We discuss confidence interval estimation of  $\bar{y}$  in two cases. In the first case where all  $x_i$  ( $i = 1, \dots, N$ ) are known, a natural estimator of  $\bar{y}$  is  $(\sum_{i \in s} y_i + \sum_{i \notin s} \hat{y}_i)/N$ . However, for the purpose of constructing confidence intervals for  $\bar{y}$ , we study the distribution of

$$\hat{\bar{y}}(\hat{\alpha}, \hat{\beta}) = \frac{1}{N} \sum_{i=1}^N \hat{y}_i = \int_{-\infty}^{\infty} h^{-1}(\hat{\alpha} + \hat{\beta} u) dF_N(u) \quad (2.4)$$

instead, where  $F_N(u)$  is the empirical distribution function of the  $u_i$  ( $i = 1, \dots, N$ ). Clearly, the distribution of  $\hat{\bar{y}}(\hat{\alpha}, \hat{\beta})$  is determined by the distribution of  $(\hat{\alpha}, \hat{\beta})$  which is described in the following design-based theorem.

**Theorem 2.1** Suppose that when  $k \rightarrow \infty$ , both  $n = n_k$  and  $N - n = N_k - n_k$  go to  $\infty$  and

1.  $\bar{u} = \lim_{k \rightarrow \infty} N^{-1} \sum_{i=1}^N u_i$  exists.
2.  $N^{-1} \sum_{i=1}^N u_i^4 = O(1)$ .
3.  $\sigma_u^2 = \lim_{k \rightarrow \infty} \sigma_{u,N}^2 = \lim_{k \rightarrow \infty} (N-1)^{-1} \sum_{i=1}^N (u_i - \bar{u}_N)^2$  exists and is greater than zero.
4.  $\sigma^2 = \lim_{k \rightarrow \infty} \sigma_N^2 = \lim_{k \rightarrow \infty} (N-1)^{-1} \sum_{i=1}^N e_i^2$  exists and is greater than zero.
5.  $N^{-1} \sum_{i=1}^N |e_i|^3 = O(1)$ ,  $N^{-1} \sum_{i=1}^N |(u_i - \bar{u}_N) e_i|^3 = O(1)$ .
6.  $r = \lim_{k \rightarrow \infty} (\sigma_{u,N}^2 \sigma_N^2)^{-1} N^{-1} \sum_{i=1}^N (u_i - \bar{u}_N)^2 e_i^2$  exists and is greater than zero.
7.  $f = \lim_{k \rightarrow \infty} n/N$  exists and is less than 1.

Then

- (1)  $\sqrt{n}(\hat{\alpha} - \alpha_N, \hat{\beta} - \beta_N)'$  converges in distribution to the bivariate normal distribution  $N_2(0, \Sigma)$ , where

$$\Sigma = \begin{pmatrix} 1 + \frac{\bar{u}^2}{\sigma_u^2} r & -\frac{\bar{u}}{\sigma_u^2} r \\ -\frac{\bar{u}}{\sigma_u^2} r & \frac{1}{\sigma_u^2} r \end{pmatrix} (1-f) \sigma^2.$$

- (2) Let  $B_n$  be any joint  $100(1 - \gamma)\%$  confidence region for  $(\alpha_N, \beta_N)$  and define  $G_n$  by

$$G_n = \{\hat{\bar{y}}(\alpha, \beta) : (\alpha, \beta) \in B_n\}, \quad (2.5)$$

then,

$$\text{Prob}\{\bar{y}(\alpha_N, \beta_N) \in G_n\} \geq 1 - \gamma,$$

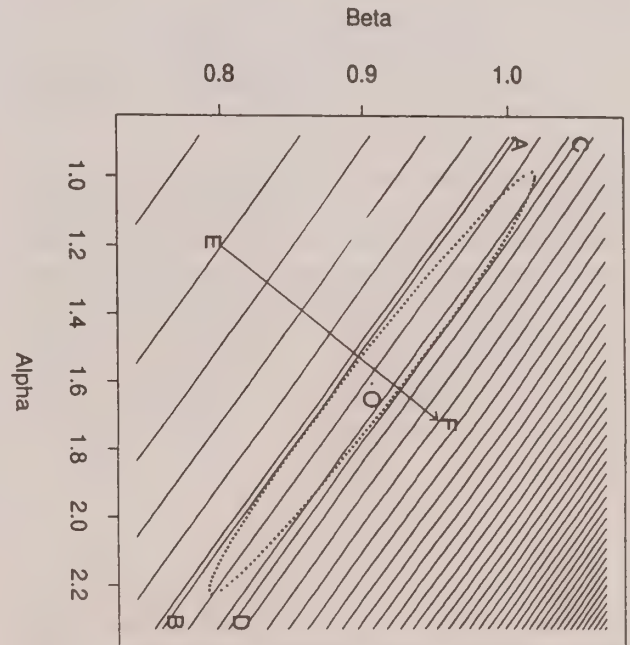
where  $\bar{y}(\alpha_N, \beta_N) = \sum_{i=1}^N h^{-1}(\alpha_N + \beta_N u_i)/N$ .

The proof is deferred to the Appendix.

We note that without underlying normality on the errors, it is not easy to get an exact confidence region  $B_n$  for  $(\alpha_N, \beta_N)$  for a specified confidence level  $1 - \gamma$ . The  $B_n$  used in the following discussion and the expressions built upon it are, therefore, approximate.

Theorem 2.1 allows us to construct confidence intervals for  $\bar{y}(\alpha_N, \beta_N)$ , but  $\bar{y}(\alpha_N, \beta_N)$  is not equal to  $\bar{y}$  in general. This is an intrinsic problem as long as a non-linear transformation is used. If only a point estimator is needed, we would use the regression estimator currently, and we suggest that the methodology developed in this paper be used for interval estimation. Bias corrections for  $\hat{\bar{y}}(\hat{\alpha}, \hat{\beta})$  are, however, possible, and a specific one is used in our simulation study. Work on general corrections is under study.

According to Theorem 2.1,  $G_n$  is a conservative confidence interval for  $\bar{y}(\alpha_N, \beta_N)$ , which can also be regarded as an approximate confidence interval for  $\bar{y}$ . To improve the coverage rate of  $G_n$ , observe that the contours of  $\hat{\bar{y}}(\alpha, \beta)$  in a small neighborhood of  $O = (\hat{\alpha}, \hat{\beta})$  are approximately parallel straight lines on the  $\alpha\beta$  plane; see Figure 1. Let  $(a, b)$  be the



**Figure 1.** Contour plot of the bi-variate function  $\hat{\bar{y}}(\alpha, \beta)$  in the neighbourhood of  $O = (\hat{\alpha}, \hat{\beta})$ , based on a random sample of size 32 taken from population Cancer

directional cosines of the direction  $\vec{EF}$  along which the contours increase. Then  $\hat{y}(\alpha, \beta)$  is (approximately) a monotone function of  $T_n = a(\alpha - \hat{\alpha}) + b(\beta - \hat{\beta})$ , where  $T_n$  is the corresponding change along the direction  $\vec{EF}$  to the changes in  $\alpha$  and  $\beta$ . A natural choice of  $B_n$  is

$$B_n = \{(\alpha, \beta) : |a(\alpha - \hat{\alpha}) + b(\beta - \hat{\beta})| \leq c\hat{\sigma}t(\gamma/2; n-2)\},$$

where  $c^2 = \text{Var}(T_n)/\sigma^2$ ,  $\text{Var}(T_n)$  is the variance of  $T_n$ , and  $t(\gamma/2; n-2)$  is the upper  $\gamma/2$ -percentile of the  $t$  distribution with  $n-2$  degrees of freedom. This  $B_n$  is the region between two parallel straight lines  $AB$  and  $CD$  in Figure 1.

A drawback of the above  $B_n$  is that it is an *unbounded* region. If the contours of  $\hat{y}(\alpha, \beta)$  are not close to be parallel and/or straight, this  $B_n$  will lead to very conservative confidence intervals. To guard against this possibility, we construct a bounded elliptic region  $C_n$  defined by those  $(\alpha, \beta)$  that satisfy

$$\begin{aligned} & \left\{ n(\alpha - \hat{\alpha})^2 + 2n\bar{u}_s(\alpha - \hat{\alpha})(\beta - \hat{\beta}) + \right. \\ & \left. n\left( \bar{u}_s^2 + r_s^{-1} \frac{\sum_{i \in s} (u_i - \bar{u}_s)^2}{n-1} \right) (\beta - \hat{\beta})^2 \right\} \\ & \leq \left( 1 - \frac{n}{N} \right) \hat{\sigma}^2 t^2(\gamma/2; n-2), \end{aligned}$$

where  $(1 - n/N)$  is part of the variances of  $\hat{\alpha}$  and  $\hat{\beta}$ , because we are doing sampling without replacement from a finite population, and

$$r_s = \frac{n^{-1} \sum_{i \in s} (u_i - \bar{u}_N)^2 (v_i - \hat{\alpha} - \hat{\beta} u_i)^2}{\left\{ n^{-1} \sum_{i \in s} (u_i - \bar{u}_N)^2 \right\} \left\{ (n-2)^{-1} \sum_{i \in s} (v_i - \hat{\alpha} - \hat{\beta} u_i)^2 \right\}} \quad (2.6)$$

is a sample estimate of the quantity  $r$  in Theorem 2.1. The  $C_n$  thus defined is represented by the region inside the ellipse in Figure 1 and has the property that it touches both boundary lines of  $B_n$  regardless of the direction  $(a, b)$ . Therefore, when  $\hat{y}(\alpha, \beta)$  is indeed a monotone function of  $T_n$ ,  $C_n$  produces the same confidence interval for  $\bar{y}$  as  $B_n$  does. However,  $C_n$  is less vulnerable than  $B_n$  if the contours of  $\hat{y}(\alpha, \beta)$  are not close to be parallel and/or straight, because  $C_n$  shrinks to one point as  $n$  increases. A confidence interval for  $\bar{y}$  corresponding to  $C_n$  is defined as

$$I_n = \{\hat{y}(\alpha, \beta) : (\alpha, \beta) \in C_n\}. \quad (2.7)$$

As the error distributions are more symmetric after the transformation, the new confidence interval based on  $C_n$  is therefore expected to be better than the confidence interval without transformation. Note that since all  $x_i$  are known, other approaches, such as optimal stratification and post-stratification, may be better. However, optimal stratification

may not be possible in some cases as discussed in Cochran (1977, p. 134). Also research is needed on the use of post-stratification when the error distributions are severely skewed.

We now turn to the discussion of the second case where  $\bar{x} = (x_1 + \dots + x_N)/N$  is known, but  $x_i$ ,  $i = 1, \dots, N$ , are unknown. If we want to proceed as in the first case, one approach is to estimate  $F_N(u)$  and somehow make use of the information in  $\bar{x}$ . The following empirical likelihood methodology is found to be an effective way of doing this. We outline the main ideas here; the interested reader should consult Owen (1988, 1990) and Chen and Qin (1992) for more details. The key idea is to maximize the (empirical) likelihood functions under various restrictions formed by the knowledge about some aspects of the parameters. For example, in our problem, the knowledge is  $\bar{x}$ . It is shown by Chen and Qin (1992) that the resulting estimators with the presence of restrictions are asymptotically more efficient than those without restrictions.

Specifically, we estimate  $F_N(u)$  in (2.4) by

$$\hat{F}_N(u) = \sum_{i \in s} p_i I[u_i \leq u], \quad (2.8)$$

where the  $p_i$  are chosen by maximizing

$$\prod_{i \in s} p_i \quad (2.9)$$

subject to

$$p_i \geq 0, \quad \sum_{i \in s} p_i = 1, \quad \sum_{i \in s} p_i x_i = \bar{x}. \quad (2.10)$$

If  $y_i$ ,  $i \in s$  are regarded as realizations of the random variables  $Y_i$ ,  $i \in s$ , with distribution function  $F$ , the  $p_i$  in (2.9) can be defined by  $p_i = F(Y_i) - F(Y_i^-)$ , and (2.9) is called the empirical likelihood function in Owen (1990).

Deville and Särndal (1992) look at the above approach from a calibration point of view. They suggest using unequal weights for different units sampled to reflect their different contributions, while keeping  $\sum p_i x_i = \bar{x}$ . It is believed that if these weights give a perfect estimate of  $\bar{x}$ , they should also be good for estimating  $\bar{y}$ .

The solution to (2.9) and (2.10) will not exist if either the minimum  $x$  value in a sample is greater than or equal to  $\bar{x}$ , or the maximum  $x$  value in a sample is less than or equal to  $\bar{x}$ . When this happens, one remedy is to replace (2.9) with

$$\sum_{i \in s} (np_i - 1)^2, \quad (2.11)$$

subject to a milder constraint

$$\sum_{i \in s} p_i = 1, \quad \sum_{i \in s} p_i x_i = \bar{x}. \quad (2.12)$$

Under (2.11) and (2.12), we have

$$p_i = \frac{1}{n} + (\bar{x} - \bar{x}_s)(x_i - \bar{x}_s) / \sum_{i \in s} (x_i - \bar{x}_s)^2, \quad (2.13)$$



which always exists unless all the  $x_i$  in the sample are the same. The latter situation corresponds to the lack of a covariate, which implies  $p_i = n^{-1}$  if  $\bar{x} = x_i$ , or the solution does not exist if  $\bar{x} \neq x_i$ . The function given in (2.11) is called the Euclidean likelihood, which is asymptotically equivalent to the empirical likelihood (2.9) (Owen 1990).

For our simulation study in Section 3, we suggest a bias correction to be used in our computation. If  $h(w) = g(w) = \log(w)$ , we suggest a corrected estimator of  $\bar{y}$  as

$$\hat{y}^*(\hat{\alpha}, \hat{\beta}) = \int_{-\infty}^{\infty} \exp\left\{\hat{\alpha} + \hat{\beta} u_i + \frac{1}{2} \hat{\sigma}^2\right\} F_N(u), \quad (2.14)$$

if all  $u_i, i = 1, \dots, N$  are known, and replace  $F_N(u)$  by  $\hat{F}_N(u)$  and  $\bar{u}_N$  in (2.6) by  $\bar{u}_s$  when only  $\bar{x}$  is known. This correction is motivated by model-based considerations under a normality assumption. Correspondingly,  $I_n$  of (2.7) is corrected as

$$I_n^* = \{\hat{y}^*(\alpha, \beta) : (\alpha, \beta) \in C_n\}. \quad (2.15)$$

When other power transformations are used, similar corrections can be made using the results in Pankratz and Dudley (1987).

### 3. APPLICATION TO SIX REAL POPULATIONS

The six real populations studied by Royall and Cumberland (1981a, 1981b, 1985) are summarized in Table 1. Attention was given to the variety in the type of data (demographic, economic, etc.), and in the logical relationship between the  $x$  and  $y$  variables, when these populations were chosen. Note that we have added 1 to the  $y$  values in population Cancer in order to take the log transformation.

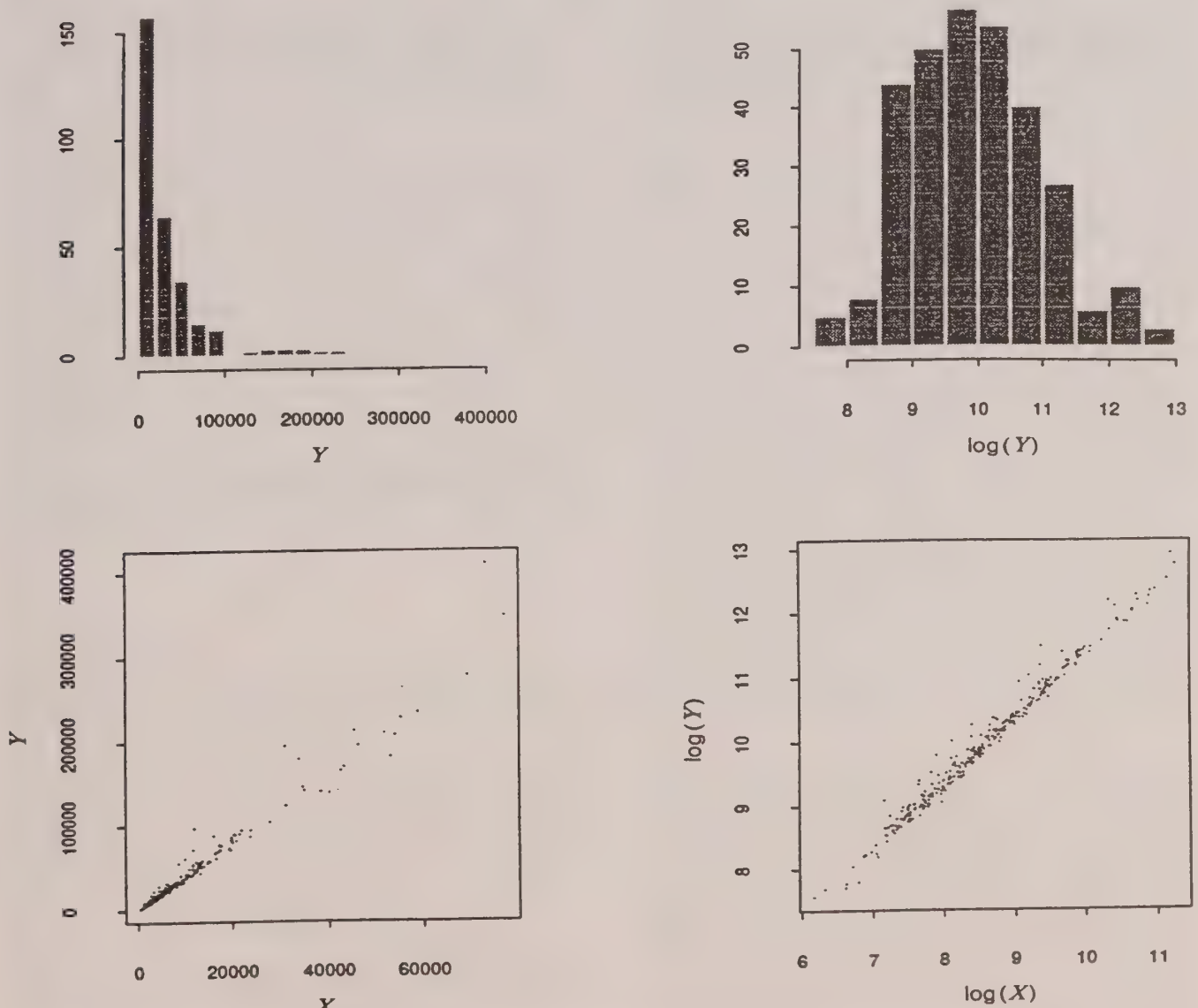


Figure 2. Histograms and scatter plots for the population Counties 70 before and after taking the log transformation

**Table 1**  
Summaries of the Six Populations

Population	$N$	$\bar{x}$	$\bar{y}$	$\rho(x, y)$	$\rho(\log(x), \log(y))$
Cancer	301	$1.1288 \times 10^4$	$4.0847 \times 10^1$	0.967	0.948
Cities	125	$2.6602 \times 10^5$	$2.8553 \times 10^5$	0.947	0.953
Counties 60	304	$8.9312 \times 10^3$	$3.2916 \times 10^4$	0.998	0.998
Counties 70	304	$8.9312 \times 10^3$	$3.6984 \times 10^4$	0.982	0.991
Hospitals	393	$2.7470 \times 10^2$	$8.1465 \times 10^2$	0.911	0.943
Sales	331	$2.3164 \times 10^9$	$2.4078 \times 10^9$	0.997	0.985

The Counties 70 data are plotted in Figure 2. The histogram of  $y$  clearly indicates that the population distribution is severely skewed, while the same plot for  $\log(y)$  shows a substantial improvement. Also, the scatter plot of  $\log(y)$  vs.  $\log(x)$  shows a better linear relationship than the scatter plot of  $y$  vs.  $x$ . The need and the benefit of taking transformation is therefore obvious. Similar comments can also be made for populations Cities, Counties 60 and Hospitals. For populations Cancer and Sales, the log transformation (or any other power transformations) seem to weaken the linear relationship that exists between  $x$  and  $y$ .

Now, we illustrate our new procedure by assuming  $h = g = \log$  in (2.2). Equations (2.9) to (2.15) are used to perform the calculations. As in Royall and Cumberland (1981b, 1985), for each of the six populations, we take a simple random sample of size 32 and calculate  $\bar{x}_s, \hat{y}^*(\hat{\alpha}, \hat{\beta})$  and construct a 95% confidence interval  $I_{32}^*$ . We repeat this process 10,000 times for each population. The results are reported in Table 2 under the title "Transformation Method" when all  $x$  values are known, and under the title "Empirical Likelihood Method" when only  $\bar{x}$  is known. The term ratio denotes the average length of the confidence intervals divided by the root mean square error for each population. The non-coverage rate (Ncr) is the proportion of intervals that fail to contain the population average  $\bar{y}$ . The quantities under the titles "Regression Method (regression variance)" and "Regression Method (jackknife variance)" are obtained using the same method of Royall and Cumberland (1981b) when the usual regression variance and the jackknife variance of  $\hat{y}$  are used, respectively, but for 10,000 random samples instead of the original 1,000 samples. The results under "Empirical Likelihood Method (created population)" are to be explained in the next Section.

Next, we follow Royall and Cumberland to make *design based* inference and to study the *conditional* coverage properties of several interval estimation procedures. Specifically, we divide the confidence intervals into 20 groups according to the size of  $\bar{x}_s$ , and plot the proportions of intervals in each group that fail to contain the population average  $\bar{y}$ . For each specific group, the proportion of those intervals that lay above (below)  $\bar{y}$  is plotted above (below) the horizontal line. Figure 3 contains such plots for the Counties 70 data. The top two plots show the non-coverage rates of the regression method using the usual regression variance and the jackknife

**Table 2**  
Simulation results based on 10,000 simple random samples of size 32

	Cancer	Cities	Counties 60	Counties 70	Hospitals	Sales
Regression Method (regression variance)						
Ratio	3.26	3.65	3.05	2.90	3.62	2.94
Ncr	0.141	0.116	0.146	0.271	0.098	0.176
Regression Method (jackknife variance)						
Ratio	4.03	3.88	4.03	3.57	3.93	3.95
Ncr	0.081	0.102	0.083	0.192	0.068	0.079
Transformation Method (all $x$ values are known)						
Ratio	5.08	4.00	3.75	3.76	4.04	5.41
Ncr	0.018	0.074	0.053	0.069	0.042	0.001
Empirical Likelihood Method (only $\bar{x}$ is known)						
Ratio	5.12	3.74	3.37	3.69	4.15	4.90
Ncr	0.017	0.082	0.081	0.082	0.037	0.006
Empirical Likelihood Method (created population)						
Ratio	3.92	3.92	3.97	3.96	3.90	3.99
Ncr	0.057	0.059	0.055	0.058	0.059	0.059

variance for  $\hat{y}$ ; the middle two plots show the non-coverage rates of our new procedure. The bottom left plot will be explained in Section 4. As can be seen clearly, our new procedure with a log transformation produces substantial improvement. For populations Cities, Counties 60 and Hospitals, our new procedure also produces some improvement (plots are not shown here). For populations Cancer and Sales, the new procedure produces very conservative results. This is likely due to the fact that the log transformation (or any power transformation) actually weakens the linear relationship between  $x$  and  $y$ .

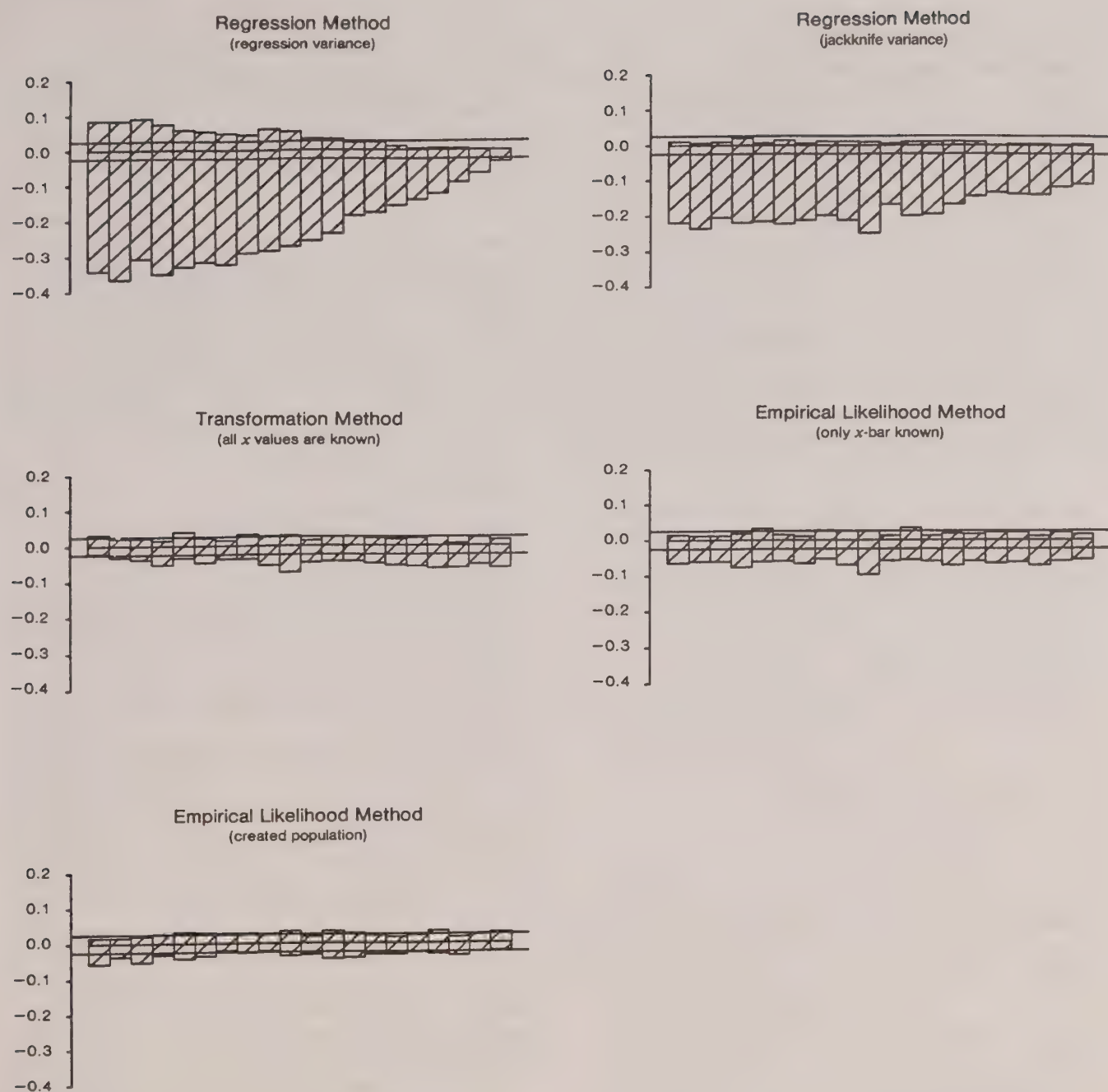
We have also performed simulations for sample sizes 16 and 64, and/or for target coverage rate 90%. The results are very similar to what we have presented.

#### 4. DISCUSSION

We use the log transformation in some of our discussions because it is perhaps the most frequently used transformation in practice. Nevertheless, there exist more objective methods to select transformations. One such a method is the well known Box-Cox power transformation which we have mentioned; see Box and Cox (1964), Box and Tidwell (1962), Carroll and Ruppert (1988). Another recent method is based on a procedure called alternating conditional expectation (ACE) (Breiman and Friedman 1985, De Veaux and Steele 1989).

There are other possibilities to improve conditional coverage rate. One such a possibility is to employ asymmetrical error distributions such as the inverse Gaussian family (Whitmore 1983). Another possibility is to adopt quasi-likelihood (Nelder and Pregibon 1987) to finite population problems.





**Figure 3.** Plots of conditional non-coverage rates for the population Counties 70 based on 10,000 simple random samples of size 32. Reference lines are drawn at 2.5% and the expected non-coverage rate is 5%

The validity of our new procedure is also demonstrated in the following simulation study. For each of the six real populations, we create a new population by replacing the original  $y_i$  values with

$$y_i^* = \exp\{\hat{\alpha} + \hat{\beta} \log(x_i) + \hat{\sigma} \epsilon_i\},$$

where  $\hat{\alpha}$ ,  $\hat{\beta}$  and  $\hat{\sigma}$  are the parameter estimates from fitting model (2.2) with  $h = g = \log$  to the old population, and  $\epsilon_i$  are generated as i.i.d. standard normal variates. Using the six

created populations which are fixed, we repeat the simulations as in Section 3 for the case where only  $\bar{x}$  is known. Table 2 contains the summary of this simulation study, and the non-coverage plot for the Counties 70 data is shown at the bottom left corner of Figure 3. (Non-coverage plots for other populations look very similar to this plot.) It is clear from this study that when the finite population is generated from a super-population model like (2.2) with a normal error distribution, our new procedure gives the correct conditional coverage rates. Furthermore, we decrease the correlation between

$x$  and  $y$  to as low as 0.5 for each of the six populations by increasing  $\hat{\sigma}$  and repeat the above simulations. The results are as good as those shown in Table 2 and Figure 3.

Although only the simple random sampling scheme is considered in this paper, the proposed procedure is applicable as long as (i) there is a linear correlation between  $h(y)$  and  $g(x)$  for some monotone functions  $h$  and  $g$ , and (ii) either  $F_N(u)$  or  $\hat{F}_N(u)$  can be found. Since the six populations studied here are carefully chosen to be representative, our new procedure is expected to be useful to study other finite populations.

## ACKNOWLEDGEMENTS

We would like to thank the referee and the Editor for their comments which greatly improved the presentation. Both authors are supported by grants from the Natural Sciences and Engineering Research Council of Canada.

## APPENDIX

**Proof of Theorem 2.1 (1).** For any given real numbers  $t_1$  and  $t_2$ , we have

$$t_1(\hat{\alpha} - \alpha_N) + t_2(\hat{\beta} - \beta_N) = t_1 n^{-1} \sum_{i \in s} e_i + \frac{t_2 - t_1 \bar{u}_s}{\sum_{i \in s} (u_i - \bar{u}_s)^2} \sum_{i \in s} (u_i - \bar{u}_s) e_i.$$

From Conditions 1, 2 and 3, we have

$$\bar{u}_s - \bar{u}, \quad n^{-1} \sum_{i \in s} (u_i - \bar{u}_s)^2 \rightarrow \sigma_u^2.$$

Therefore, we can write

$$t_1(\hat{\alpha} - \alpha_N) + t_2(\hat{\beta} - \beta_N) = t_1 n^{-1} \sum_{i \in s} e_i + \frac{t_2 - t_1 \bar{u}}{\sigma_u^2} n^{-1} \sum_{i \in s} (u_i - \bar{u}) e_i + o_p(n^{-1/2}).$$

The Lindeberg-Hájek condition is satisfied for  $t_1 e_i + t_2 - t_1 \bar{u} / \sigma_u^2 (u_i - \bar{u}) e_i$  under the moment condition 5, see Hájek (1960), Scott and Wu (1981) and Bickel and Freedman (1984). Together with Conditions 4, 6 and 7, the desired result follows by using the Cramér-Wold device.

**Proof of Theorem 2.1 (2).** Because there may be other values  $(\alpha', \beta') \notin B_n$  for which  $\hat{y}(\alpha', \beta') = \hat{y}(\alpha, \beta)$  for some  $(\alpha, \beta) \in B_n$ ,  $G_n$  is always conservative.

## REFERENCES

BICKEL, P.J., and FREEDMAN, D.A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *The Annals of Statistics*, 12, 470-482.

- BOX, G.E.P., and TIDWELL, P.W. (1962). Transformations of the independent variables. *Technometrics*, 4, 531-550.
- BOX, G.E.P., and COX, D.R. (1964). An analysis of transformations. *Journal of The Royal Statistical Society, Series B*, 26, 211-243, discussions 244-252.
- BREIMAN, L., and FRIEDMAN, J. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80, 580-597.
- CARROLL, R.J., and RUPPERT, D. (1981). On prediction and power transformation family. *Biometrika*, 68, 609-615.
- CARROLL, R.J., and RUPPERT, D. (1988). *Transformation and Weighting in Regression*. London: Chapman and Hall.
- CALVIN, J.A., and SEDRANSK, J. (1991). Bayesian and frequentist predictive inference for the patterns of care studies. *Journal of the American Statistical Association*, 86, 36-54.
- CHEN, J., and QIN, J. (1992). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika*, 80, 107-116.
- COCHRAN, W.G. (1977). *Sampling Techniques*. (3rd ed.) New York: John Wiley.
- DE VEAUX, R.D., and STEELE, J.M. (1989). ACE guided-transformation method for estimation of the coefficient of soil-water diffusivity. *Technometrics*, 31, 91-98.
- DEVILLE, J., and SÄRNDAL, C. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- ELLIOTT, J.M. (1977). Statistical analysis of samples of benthic invertebrates. *Freshwater Biological Scientific Publication*, No. 25, (2nd ed.).
- HÁJEK, J. (1960). Limiting Distributions in Simple Random Sampling From a Finite Population. *Publications in Mathematics of the Hungarian Academy of Science*, 5, 361-374.
- NELDER, J.A., and PREGIBON, D. (1987). An extended quasi-likelihood function. *Biometrika*, 74, 221-232.
- OWEN, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75, 237-249.
- OWEN, A. (1990). Empirical likelihood confidence regions. *The Annals of Statistics*, 18, 90-120.
- PANKRATZ, A., and DUDLEY, U. (1987). Forecasts of power-transformed series. *Journal of Forecasting*, 6, 239-248.
- ROYALL, R.M., and CUMBERLAND, W.G. (1981a). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, 76, 66-88.
- ROYALL, R.M., and CUMBERLAND, W.G. (1981b). The finite-population linear regression estimator and estimators of its variance - An empirical study. *Journal of the American Statistical Association*, 76, 924-930.
- ROYALL, R.M., and CUMBERLAND, W.G. (1985). Conditional coverage properties of finite population confidence intervals. *Journal of the American Statistical Association*, 80, 355-359.
- SÄRNDAL, C., SWENSSON, B., and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- SCOTT, A., and WU, C.F. (1981). On the asymptotic distribution of the ratio and regression estimators. *Journal of the American Statistical Association*, 76, 98-102.
- WHITMORE, G.A. (1983). A regression method for censored inverse-gaussian data. *The Canadian Journal of Statistics*, 11, 305-315.



# The Application of McNemar Tests to the Current Population Survey's Split Panel Study

KATHERINE JENNY THOMPSON and ROBIN FISHER<sup>1</sup>

## ABSTRACT

Results from the Current Population Survey split panel studies indicated a centralized computer-assisted telephone interviewing (CATI) effect on labor force estimates. One hypothesis is that the CATI interviewing increased the probability of respondent's changing their reported labor force status. The two sample McNemar test is appropriate for testing this type of hypothesis: the hypothesis of interest is that the marginal changes in each of two independent sample's tables are equal. We show two adaptations of this test to complex survey data, along with applications from the Current Population Survey's Parallel Survey split panel data and from the Current Population Survey's CATI Phase-in data.

**KEY WORDS:** Current Population Survey; Parallel survey; Nonparametric statistics.

## 1. INTRODUCTION

Results from the Current Population Survey's Parallel Survey split panel study and from the Current Population Survey's CATI Phase-in Project provided some indication of a centralized computer-assisted telephone interviewing (CATI) effect on the United States' monthly labor force estimates (Thompson 1994 and Shoemaker 1993). One hypothesis is that the CATI interviewing increased the probability of respondent's changing their reported labor force status from the first (personal) interview to the second (CATI) interview.

The two sample McNemar test is appropriate for testing this type of hypothesis. The McNemar test (1947) has been generalized to a two sample situation where the hypothesis of interest is that the marginal changes in each of two independent samples'  $2 \times 2$  tables are equal (Feuer and Kessler 1989). The application presented was for a two sample cohort analysis and assumed simple random sampling.

Certain modifications of the test statistic for a McNemar test are necessary for a complex survey data application. First, because the data are not obtained through a simple random sample and are weighted, a separate estimate of the variance is required. Second, unless the survey has a longitudinal design, a separate link of individuals in two consecutive months' of data must be performed. In general, such a link will include some false matches and exclude some true matches. This adds another source of variance.

We show two adaptations of this test to complex survey data. In particular, we present these tests along with applications to the Current Population Survey's Parallel Survey split panel study and from the Current Population Survey's CATI Phase-in Project. In Section 2 we describe these test modifications including background on the one and two-sample McNemar tests (Section 2.1), modifications for

complex survey data (Section 2.2), and some remarks on applications to several months' data (Section 2.3). Section 3 presents applications of these tests specifically to Current Population Survey Parallel Survey Data and to Current Population Survey CATI Phase-in data including background on the two studies (Section 3.1), details of the panel estimates and variance estimates (Section 3.2), diagnostics (Section 3.3), and results (Section 3.4). We make some concluding remarks in Section 4. Details of covariance estimation are included in the appendix.

## 2. TEST AND MODIFICATIONS

### 2.1 General

A sample is randomly split into two independent representative samples (split panels). After a baseline measurement is taken in both panels, a new technique is administered in one panel, the treatment panel. The other panel serves as a control.

The records are linked longitudinally after the second measured. A matched response can be +, -, or \* (missing). Since this is matched data, the "\*\*\*" cell will be empty.

This scenario is represented pictorially as

		Treatment Panel			
		Month 2			
		Treatment			
		+   -   *			
Month 1	+	$x_{++}$	$x_{-+}$	$x_{*+}$	$x_{+.$
	-	$x_{+-}$	$x_{--}$	$x_{*-}$	$x_{-.$
	*	$x_{*+}$	$x_{*-}$		$x_{*.}$
		$x_{.+}$	$x_{.-}$	$x_{.*}$	$n$

<sup>1</sup> Katherine Jenny Thompson, Economic Statistical Methods and Programming Division, and Robin Fisher, Housing and Household Economic Statistics Division, United States Bureau of the Census, Washington, DC 20233, U.S.A.

Control Panel		Month 2			
		No Treatment			
		+	-	*	
Month 1	+	$x'_{++}$	$x'_{+-}$	$x'_{+*}$	$x'_{+}$
No Treatment	-	$x'_{-+}$	$x'_{--}$	$x'_{-*}$	$x'_{-}$
	*	$x'_{*+}$	$x'_{*-}$		$x'_{*}$
		$x'_{+}$	$x'_{-}$	$x'_{*}$	$n'$

where  $n$  is not necessarily equal to  $n'$ .

For each panel, define

$M_{(12)}$  as the set of cases which have month 1 and month 2 responses (matched cases). This set contains  $n_{(12)} = (x_{++} + x_{+-} + x_{-+} + x_{--})$  elements;

$M_{(10)}$  as the set of cases which have month 1 responses, but no month 2 response. This set contains  $n_{(10)} = (x_{+*} + x_{-*})$  elements;

$M_{(02)}$  as the set of cases which have month 2 responses, but no month 1 response. This set contains  $n_{(02)} = (x_{*+} + x_{*-})$  elements.

Note that the  $n$ 's are sample sizes and do not have weights.

First, consider the one-sample case. Traditionally, the one-sample McNemar test statistic is constructed from the  $n_{(12)}$  and  $n'_{(12)}$  matched responses, where a prime (') indicates the control panel. In the one-sample scenario, we test the hypothesis

$$H_0: p_{+-} = p_{-+}, \text{ where the } p\text{'s refer to cell probabilities}$$

$$H_1: \text{Not } H_0$$

i.e., the hypothesis that the movement from one state to the other (+ to -, or - to +) is zero. We also refer to this movement as the flux.

The one-sample test can be a useful diagnostic in the two-sample situation. We examine the Control panel estimates to see if there is zero movement. Any significant movement in the Treatment panel can be measured as a deviation from zero flux or as a change in the probability of a "+."

The two-sample hypothesis is

$$H_0: (p_{+-} - p_{-+}) = (p'_{+-} - p'_{-+})$$

$$H_1: \text{Not } H_0$$

In other words, the difference in the probabilities of switching in the two directions is the same, regardless of the treatment, or equivalently, the difference in panel fluxes is zero.

The Feuer and Kessler generalization (1989) to a two-sample McNemar test (described in 2.2.1 below) is confined to the  $M_{(12)}$  and  $M'_{(12)}$  linked sets. We can add an additional assumption, however, to allow the unmatched responses to be included in computation of the test statistics. This assumption motivates the discussion in Section 2.2.2.

## 2.2 Complex Survey Modifications

### 2.2.1 Modification One: Longitudinally Linked Data

This method is a straightforward application of the two-sample McNemar test, using longitudinally linked data from a complex survey.

To construct the test statistic, we examine the cell probabilities and note that

$$[p_{-+} - p_{+-}] = [(p_{++} + p_{-+}) - (p_{++} + p_{-+})]$$

$$= [p_{+}^* - p_{-}^*]$$

$$= p_2^* - p_1^*$$

where  $p_2^*$  is the marginal probability of a + response month 2, given a matched response for both months; and  $p_1^*$  is the marginal probability of a + response month 1, given a matched response for both months.

The one-sample test statistic constructed from this panel's data is

$$Z_1^* = \frac{p_2^{\circ} - p_1^{\circ}}{\sqrt{\text{Var}(p_2^{\circ} - p_1^{\circ})}},$$

where

$$p_1^{\circ} = \frac{x_{++} + x_{+-}}{n_{(12)}}, \quad p_2^{\circ} = \frac{x_{++} + x_{-+}}{n_{(12)}}.$$

Given two independent panels, the two-sample test statistic is

$$Z^* = \frac{(p_2^{\circ} - p_1^{\circ}) - (p_2'^{\circ} - p_1'^{\circ})}{\sqrt{\text{Var}(p_2^{\circ} - p_1^{\circ}) + \text{Var}(p_2'^{\circ} - p_1'^{\circ})}},$$

where

$$p_1'^{\circ} = \frac{x'_{++} + x'_{+-}}{n'_{(12)}}, \quad p_2'^{\circ} = \frac{x'_{++} + x'_{-+}}{n'_{(12)}}.$$

These results hold regardless of sample design. To extend the results to a complex survey application, we use weighted estimates and complex survey variances and covariances in place of simple random sample variances.

If the survey is designed to collect longitudinal data, then this modification is a natural extension of the method described by Feuer and Kessler. For this type of survey design, an effective mechanism to link individuals from month to month is presumably in place. Often, however, this is not the case, and one data set must be physically linked to another. Consequently, the  $n_{(12)}$  elements in the domain will contain some false matches, and some actual matches may be inadvertently excluded. Both the record weights and variance estimates will need to be adjusted to account for the matching. Jabine and Scheuren (1986) provide an excellent summary of the methodological issues arising from the use of linked data, both for model-based and ad-hoc ("hard") record linkage techniques.



### 2.2.2 Modification Two: Unlinked Data

This method omits the longitudinal linkage step altogether, noting that the construction of the traditional McNemar test statistic can be expressed in terms of estimates of marginal probabilities. Assume that under the null hypothesis, the expected value of  $(p_{*+} - p_{+*})$  is zero. This is described for a simple random sampling application in Marascuilo *et al.* (1988).

The domain for the first month of data is given  $M_{(12)} \cup M_{(10)}$  which contains  $n_{(12)} + n_{(10)} = n_1$  elements. The domain for the second month of data is given by  $M_{(12)} \cup M_{(02)}$  which contains  $n_{(12)} + n_{(02)} = n_2$  elements.

The one-sample test statistic constructed from the unlinked data is given by

$$Z_1 = \frac{p_2 - p_1}{\sqrt{\text{Var}(p_2 - p_1)}},$$

where

$$p_1 = \frac{x_{*+}}{n_1}, \quad p_2 = \frac{x_{+*}}{n_2}.$$

Given two independent panels, the two-sample test statistic is

$$Z = \frac{(p_2 - p_1) - (p'_2 - p'_1)}{\sqrt{\text{Var}(p_2 - p_1) + \text{Var}(p'_2 - p'_1)}},$$

where

$$p'_1 = \frac{x'_{*+}}{n'_1}, \quad p'_2 = \frac{x'_{+*}}{n'_2}.$$

As with the application described in 2.2.1, all estimates are weighted estimates, and variances are complex survey variances.

### 2.3 Linear Combinations

We can use our estimated covariance matrix to test linear combinations of  $\hat{\lambda}_T$ ,  $\hat{\lambda}_C$ , and  $\hat{\delta}$  over time, where  $\hat{\lambda}_T = p_2 - p_1$ ,  $\hat{\lambda}_C = p'_2 - p'_1$ , and  $\hat{\delta} = \hat{\lambda}_T - \hat{\lambda}_C$ , and  $p_1, p_2, p'_1$  and  $p'_2$  are vectors containing the marginal probabilities for the time period under consideration.

General linear hypotheses of the form  $K'\mu$  are now easily tested. One might wish to test for contrast by time period, for example testing the average difference from January through June against the remainder of the year's data. Perhaps the most interesting (to our applications) of these tests is of the hypothesis  $H_0: \underline{1}'\mu = 0$ , where  $\mu$  is the expected value of one of the vectors described above.

Another test of particular interest is the "omnibus hypothesis," where we test  $H_0: \mu = 0$ . The test statistics for this hypothesis are  $\hat{\lambda}_T' \sum_{\lambda(T)}^{-1} \hat{\lambda}_T$ ,  $\hat{\lambda}_C' \sum_{\lambda(C)}^{-1} \hat{\lambda}_C$ , and  $\hat{\lambda}_\delta' \sum_{\lambda(\delta)}^{-1} \hat{\lambda}_\delta$ , each of which has an approximate chi-squared distribution with  $r$  degrees of freedom, where  $r$  is the dimension of the vector of interest.

## 3. APPLICATIONS

In this section, we apply the one and two-sample McNemar techniques for unlinked data outlined in 2.2.2 and 2.3 to two separate sets of data: the Current Population Survey's Parallel Survey split panel data and Current Population Survey CATI Phase-in data. Tables 1 and 2 (section 3.4.1) provide the results for Parallel Survey split panel data. Tables 3 and 4 (section 3.4.2) provide the results for the Current Population Survey CATI Phase-in data.

### 3.1 Background

The official monthly civilian labor force estimates from January 1994 onward are based on data from a comprehensively redesigned Current Population Survey. The redesign included implementation of a new, fully computerized questionnaire, and an increase in centralized computer-assisted telephone interviewing (CATI). To gauge the effect of the Current Population Survey redesign on published estimates, a Parallel Survey was conducted using the new questionnaire and data collection procedures from July 1992 through December 1993. Special studies were embedded in both the Parallel Survey and the Current Population Survey during the same time period to provide data for testing hypotheses about the effects of the new methodological differences on labor force estimates: the Parallel Survey split panel study and the Current Population Survey CATI Phase-in Project (a continuation of the study presented in Shoemaker 1993).

The effect of increased centralized computer-assisted telephone interviewing was of particular interest. Findings from the study described in Shoemaker (1993) had shown that including centralized telephone interviews tended to yield a larger unemployment rate. The two-sample McNemar test appeared to be a good vehicle for examining this phenomenon. In both the Current Population Survey and the Parallel Survey, households are interviewed for 4 consecutive months, not interviewed for the next 8 consecutive months, and then interviewed for another 4 consecutive months. The first and fifth interviews are conducted by a personal visit, and the subsequent interviews are conducted by telephone whenever possible. Thus the first and fifth interviews provide a baseline measurement of labor force status; the second and sixth interviews provide a "post-treatment" measurement of labor force status.

To create the panels for both studies, sample within selected sample areas was randomly divided into two representative panels using systematic sampling methods. The treatment panel was designated as CATI eligible. This meant that the sample households in the panel were eligible for interview at a centralized facility after the initial (first and fifth) interviews. To be interviewed by CATI, a respondent must have a telephone and speak English or Spanish, and must agree to be interviewed in subsequent months by telephone. Not all households in this panel were interviewed by CATI. The other panel served as a control.

The monthly unemployment rate is the primary statistic of interest published from Current Population Survey data. This rate is defined as the estimated number of unemployed persons divided by the estimated number of persons in the civilian labor force (the denominator does not include military personnel, persons under sixteen years old, or people who are no longer looking for work, or retired persons). Our primary goal was to understand how including CATI interviews influenced the probability of changing labor force status, in this case from unemployed to not unemployed (or vice versa). Our statistics for the one and two-sample McNemar tests used unemployment to population ratios, rather than unemployment rates. This allowed for a slightly more precise estimate of the proportion by decreasing the variability of the test statistic.

### 3.2 Estimates

Each month/panel estimate is an unbiased estimate. That is, the weights used to produce the estimates were strictly a function of the probability of selection: each weight is the product of the baseweight (the inverse probability of selection for a PSU), the weighting control factor (an adjustment for field subsampling), and a split panel factor (an adjustment for the probability of inclusion in a split panel). The split panel factor for the Parallel Survey study was constant by design: nine tenths of the sample was randomly assigned to the treatment panel. The split panel factors for the CPS CATI Phase-in were not constant: the sample in the treatment panel varied on a monthly level, as more sample was randomly assigned to CATI facilities.

Variances of levels were computed with generalized variance functions (GVFs). For more details, see Fisher *et al.* (1993). Robert Fay used his VPLX software (Fay 1990) to calculate replicate estimates of correlation between rotation groups for unemployed and for civilian labor force using September 1992 through December 1993 data from the Current Population Survey. We used these correlations for the test statistics based on unlinked data, assuming that they would not differ by survey (Current Population Survey versus Parallel Survey) or by geography (national versus sub-national). We derived an expression for the within-panel correlation for civilian population by relating previously calculated autocorrelations (Fisher and McGuinness 1993) and variance estimates to the individual rotation group estimates. See the appendix for details of the estimation of the correlations.

We did not use the linked modification in our applications for several reasons. The primary reason was the difficulty of longitudinally matching the data. Moreover, we were unable to evaluate the success of our matching. Finally, we did not have any estimates of correlation for the linked data.

Implicit in our analysis of the unlinked data is the assumption that the probability of a nonresponse (or a non-match) is random. We assume that the probability of a nonresponse one month is independent of the respondent's

labor force classification in the previous month. This assumption is not universally recognized. In fact, Stasny and Fienberg (1984) argue the reverse, and propose several alternative discrete-time models for the use of longitudinally linked CPS data. In our application, the estimates of marginal probabilities based on our (perhaps) poorly matched linked data were almost identical to the estimates based on unlinked data, and so we feel that our analysis did not suffer particularly from our assumption.

### 3.3 Diagnostics

Small expected sample sizes in individual cells will result in highly variable and consequently unreliable tests. We are not aware of a general method of calculating adequate sample sizes for this type of analysis using complex survey data. Instead, as a naive approach we used a slightly modified version of the traditional Pearson chi-squared test diagnostic to form a cut-off value as follows:

As defined in Section 2.2.2, let

- $x_{+ \cdot}$  = unweighted unemployed persons in month 1;
- $x_{- \cdot}$  = unweighted not-unemployed persons in month 1;
- $x_{+ \cdot}$  = unweighted unemployed persons in month 2;
- $x_{- \cdot}$  = unweighted not-unemployed persons in month 2.

Recall that in the case of the usual contingency table,  $E[+-] = x_{+ \cdot} x_{- \cdot} / n_{(12)}$ ,  $E[-+] = x_{- \cdot} x_{+ \cdot} / n_{(12)}$  under the assumption of independence (and ignoring missing values). In our estimates of expected cell size, we used unlinked marginal data. The sample sizes for the two margins corresponding to the two months are different; that is, the denominators of the expected cell totals are different depending on which margin we examine. Because we could not observe  $n_{(12)}$ , we estimated it by the geometric mean of  $n_1$  and  $n_2$ , which seemed to most closely resemble the expression for the expected cell size. We have not evaluated the effectiveness of the geometric mean versus alternative estimators.

A commonly used rule in contingency table analysis is that expected cell sizes should be at least five. However, both the Current Population Survey and Parallel Survey designs are highly clustered, and we felt that the cut-off value should be adjusted upwards. Accordingly, we multiplied the cut-off value by a design effect. We further increased the cut-off value for expected cell sizes to compensate for the correlation between the rows and columns of our tables to arrive at our final cut-off expected cell size of ten.

### 3.4 Results

#### 3.4.1 Parallel Survey Split Panel Study

This section presents the formal results from the one and two-sample McNemar tests using unlinked Parallel Survey split panel data. Although this data was collected monthly, small expected cell sizes in the control panel led us to omit several sets of adjacent months from this analysis. Table 1



**Table 1**  
One-Sample McNemar Tests for Individual Parallel  
Survey Panels – Unlinked Data

Time Frame	Treatment Panel			
	$p_2 - p_1$	$se(p_2 - p_1)$	Z-Statistic	P-Value
10/92 – 11/92	-0.62	0.29	-2.18	0.03
11/92 – 12/92	-0.47	0.28	-1.68	0.09
04/93 – 05/93	-0.76	0.27	-2.84	0.00
06/93 – 07/93	-0.04	0.27	-0.16	0.88
08/93 – 09/93	-0.66	0.27	-2.42	0.02
	Control Panel			
	$p'_2 - p'_1$	$se(p'_2 - p'_1)$	Z-Statistic	P-Value
10/92 – 11/92	2.44	0.81	3.02	0.00
11/92 – 12/92	0.11	0.83	0.14	0.89
04/93 – 05/93	0.20	0.72	0.27	0.78
06/93 – 07/93	0.97	0.71	1.38	0.17
08/93 – 09/93	-1.73	0.68	-2.54	0.01

provides summary statistics for the one-sample “monthly” tests for each panel which were based on unlinked data from the Parallel Survey’s split panels. Table 2 provides summary statistics for the two-sample tests based on unlinked data.

The reported values of  $p_1$ ,  $p_2$ ,  $p'_1$ , and  $p'_2$  are percentages of estimated unemployed to estimated total population for the panel. Recall that  $p_1$  and  $p'_1$  are the panel ratio of estimated unemployed from the first and fifth interviews to the estimated panel population from the first and fifth interviews;  $p_2$  and  $p'_2$  are the panel ratio of estimated unemployed from the second and sixth interviews to the estimated panel population from the second and sixth interviews. Data from the time frame of February 1993 – March 1993 are omitted: a CATI facility was closed during the March interview week because of a blizzard.

The one-sample McNemar tests in Table 1 test the probability that the proportion unemployed does not change between the initial and the subsequent interview within the same panel. We use the Control panel to examine the unemployment flux from one month to the next in the absence of CATI. Note that the two significant point estimates are in the opposite direction.

The entire vector of differences of proportions was found to be significantly different from the zero vector ( $p$ -value = 0.00), but the sum of the individual components was not found to be significant ( $p$ -value = 0.24). Consequently, we did not test any further linear combinations.

We expected a certain amount of month-in-sample bias to be present in these estimates. In Adams (Bureau of the Census 1991), the estimates of  $p_1$  constructed from the first and fifth months in sample of the full Current Population Survey were roughly six percent larger than their respective second and sixth month-in-sample analogues ( $p_2$ ). Consequently, estimates of  $(p_2 - p_1)$  calculated from the full Current Population Survey data were generally negative. As seen in

Table 1, this was not the case with the Parallel Survey Control panel’s estimates: counter to our intuition, the estimated difference  $(p'_2 - p'_1)$  is generally positive. This could be a function of the time difference, a geographic difference, or a design difference. Adams used 1987 data from the Current Population Survey to calculate national estimates of biases associated with rotation groups. Thus in each of these one-sample tests, the net movements are intertwined with an unmeasured effect from month-in-sample bias.

Note the negative unemployment flux in the Treatment panel. This observation is supported by the significant result from the formal test of the omnibus hypothesis ( $p$ -value = 0.00), and the significant result for the hypothesis  $\mathbf{1}'\boldsymbol{\mu} = 0$  ( $p$ -value = 0.00).

The two-sample McNemar test results are presented below.

**Table 2**  
Two-Sample McNemar Tests – Unlinked Parallel Survey Data

Time Frame	$(p_2 - p_1) - (p'_2 - p'_1)$	$se[(p_2 - p_1) - (p'_2 - p'_1)]$	Z-Statistic	P-Value
10/92 – 11/92	-3.06	0.86	-3.58	0.00
11/92 – 12/92	-0.58	0.88	-0.66	0.51
04/93 – 05/93	-0.95	0.77	-1.24	0.22
06/93 – 07/93	-1.02	0.76	-1.34	0.18
08/93 – 09/93	1.08	0.74	1.47	0.14

Individually, the monthly results do not demonstrate a clear difference in the unemployment flux between the two panels. On the other hand, the omnibus test statistic is significant ( $p$ -value = 0.00). The mean unemployment flux seems to be lower in the treatment panel as evidenced by the significant test results of the hypothesis  $\mathbf{1}'\boldsymbol{\mu} = 0$ , where  $\boldsymbol{\mu}$  is the vector of  $((p_2 - p_1) - (p'_2 - p'_1))_i$ ’s, with each element corresponding to a month’s estimate ( $p$ -value = 0.01).

In these tests, we make statements about contrasts in a table of probabilities, looking for indicators of the effect of a treatment on unemployment movement. As mentioned earlier, some month-in-sample bias is present in the one-sample tests. The tested hypotheses examine combinations of the net movement within a panel and month-in-sample bias. This problem is somewhat mitigated in the two-sample tests. Indeed, if month-in-sample bias is an additive term which affects both panels equally, it will cancel out of the test statistic. Moreover, this effect will be alleviated somewhat in the two-sample test even if it is not the same between the two panels or is multiplicative. Our preliminary sensitivity analysis bore this out: we found that the one-sample tests were sensitive to month-in-sample bias, but that the two-sample tests were not.

The two-sample  $t$ -tests presented in Thompson (1994) failed to detect a difference by panel in mean unemployment rate using the Parallel Survey split panel data. This contrasts with the Current Population Survey CATI Phase-in results: over two years, the CATI (Treatment) panel had consistently significantly higher unemployment rates than the non-CATI

(Control) panel. See Shoemaker (1993). In this analysis of Parallel Survey split panel data, we have evidence that the expected value of the proportion unemployed is lower in the presence of CATI. There are, however, some problems with the data. First, as previously mentioned, there is some confounding in the Treatment (CATI) panel, since not all respondents in this panel have their second interview conducted from a centralized telephone facility. Second, in each month the expected sample size in the Control panel cells was near ten, which could be small enough to make the distribution behave unpredictably. This latter problem is not an issue with the Current Population Survey CATI Phase-in study analysis presented in 3.4.2.

### 3.4.2 Current Population Survey CATI Phase-in Project Results

The Current Population Survey CATI Phase-in project was a continuation of the study presented in Shoemaker (1993). The primary purpose of this study was to measure the effect of including CATI interviewing on the unemployment rate. CATI interviewers in this study used an automated version of the old Current Population Survey paper questionnaire, which had a slightly modified version of the lead-in labor force question. More details are provided in Thompson (1994). The data considered in this paper are from the same time period as the Parallel Survey split panel data examined in 3.4.1: October 1992 through December 1993, again omitting the February 1993 – March 1993 time frame. Expected cell sizes in both the Treatment (CATI) and Control (non-CATI) panels were well over one hundred, and so all other contiguous months of data are included.

The one-sample McNemar test results for both panels are presented in Table 3. Test statistics are constructed with unlinked data. The reported values of  $p_1$ ,  $p_2$ ,  $p_1'$ , and  $p_2'$  are percentages of estimated unemployed to estimated total population for the panel.

As with the Parallel Survey split panel data, the one-sample McNemar tests using the CATI Phase-in data test the probability that the proportion unemployed does not change between the initial and the subsequent interview within the same panel. Again, we use the Control panel to estimate the unemployment flux from one month to the next in the absence of CATI. The monthly tests for the Control panel do not appear to exhibit any particular movement. Furthermore, the omnibus hypothesis test was not significant ( $p$ -value = 0.29), so we did not test any further linear combinations.

Again basing our expectations on the effects of month-in-sample bias presented in Adams (1991), we believed that the Control panel estimate of  $p_1'$  (from the first and fifth months-in-sample) would be larger than its respective second and sixth month-in-sample analog,  $p_2'$ . On the average, this was the case: although quite variable, the estimates of  $p_1'$  are on the average about 4 percent larger than the estimates of  $p_2'$ . Because both panels are representative samples from the same parent sample, we assume that the month-in-sample bias

**Table 3**  
One-Sample McNemar Tests for Individual Current  
Population Survey Panels – Unlinked Data

Time Frame	Treatment Panel			
	$p_2 - p_1$	$se(p_2 - p_1)$	Z-Statistic	P-Value
10/92 – 11/92	1.13	0.16	7.63	0.00
11/92 – 12/92	0.07	0.17	0.44	0.66
12/92 – 01/93	0.43	0.13	3.46	0.00
01/93 – 02/93	0.00	0.14	0.03	0.97
03/93 – 04/93	-0.25	0.14	-1.81	0.07
04/93 – 05/93	0.63	0.13	4.99	0.00
05/93 – 06/93	0.88	0.13	6.56	0.00
06/93 – 07/93	0.84	0.13	6.49	0.00
07/93 – 08/93	-0.07	0.14	-0.51	0.61
08/93 – 09/93	0.42	0.13	3.17	0.00
09/93 – 10/93	0.06	0.12	0.52	0.60
10/93 – 11/93	1.05	0.12	8.45	0.00
11/93 – 12/93	0.18	0.14	1.27	0.20
	Control Panel			
	$p_2' - p_1'$	$se(p_2' - p_1')$	Z-Statistic	P-Value
10/92 – 11/92	0.05	0.47	0.11	0.92
11/92 – 12/92	-0.14	0.47	-0.30	0.76
12/92 – 01/93	0.72	0.43	1.68	0.09
01/93 – 02/93	-0.91	0.43	-2.11	0.03
03/93 – 04/93	-0.16	0.39	-0.40	0.69
04/93 – 05/93	-0.18	0.43	-0.42	0.67
05/93 – 06/93	0.47	0.38	1.22	0.22
06/93 – 07/93	-0.32	0.46	-0.68	0.49
07/93 – 08/93	-0.52	0.39	-1.32	0.19
08/93 – 09/93	-0.54	0.44	-1.21	0.23
09/93 – 10/93	-0.08	0.37	-0.22	0.83
10/93 – 11/93	-0.63	0.42	-1.50	0.13
11/93 – 12/93	-0.09	0.37	-0.23	0.82

behaves similarly in both panels. The Treatment (CATI) panel estimates of  $p_2$  are *larger* on the average than the estimates of  $p_1$ . Given the Control panel's estimates behavior, this phenomenon provides some evidence of a CATI effect.

Note the movement in the Treatment panel from *not* unemployed to unemployed. This observation is supported by the significant result from the formal test of the omnibus hypothesis ( $p$ -value = 0.00), and the significant result for the hypothesis  $1'\mu = 0$  ( $p$ -value = 0.00). In contrast to the Parallel Survey results provided in 3.4.1, this data provides some evidence that unemployment rate is higher in the presence of CATI. This evidence is further supported by the two sample McNemar test results provided Table 4. The individual monthly results in Table 4 provide some evidence of difference in the unemployment flux between two panels. Furthermore, the omnibus test is significant ( $p$ -value = 0.00). The mean unemployment flux in the Treatment panel seems to be higher as evidenced by the significant test results of the hypothesis  $1'\mu = 0$ .

The two-sample  $t$ -tests presented in Thompson (1994) also detected a *positive* difference by panel in mean unemployment rate using the Current Population Survey split panel data



**Table 4**  
Two-Sample McNemar Tests – Unlinked Current  
Population Survey Data

Time Frame	$(p_2 - p_1) - (p_2' - p_1')$	$se[(p_2 - p_1) - (p_2' - p_1')]$	Z-Statistic	P-Value
10/92 – 11/92	1.18	0.50	2.38	0.02
11/92 – 12/92	0.22	0.50	0.43	0.67
12/92 – 01/93	-0.29	0.45	-0.64	0.52
01/93 – 02/93	0.92	0.45	2.03	0.04
03/93 – 04/93	-0.10	0.42	-0.23	0.81
04/93 – 05/93	0.81	0.45	1.81	0.07
05/93 – 06/93	0.41	0.41	1.01	0.31
06/93 – 07/93	1.16	0.48	2.41	0.02
07/93 – 08/93	0.45	0.42	1.07	0.28
08/93 – 09/93	0.95	0.46	2.06	0.04
09/93 – 10/93	0.14	0.39	0.37	0.71
10/93 – 11/93	1.69	0.44	3.83	0.00
11/93 – 12/93	0.26	0.40	0.66	0.51

*i.e.*, including CATI interviews resulted in a *higher* unemployment rate. These results were consistent with the Current Population Survey CATI Phase-in results presented in Shoemaker (1993). This analysis of Current Population Survey split panel data reinforces that conclusion. Again, it is impossible to attribute the positive net migration from not unemployed to unemployed entirely to the effect of CATI: the same confounding described in 3.4.1 is present in this Treatment (CATI) panel.

### 3.5 Discussion

Our results appear to yield opposite conclusions about the effect of CATI on unemployment flux. The CATI effect is not, however, the same in both tests.

Perhaps the key difference is the questionnaire. The Parallel Survey data was collected using the newly redesigned Current Population Survey questionnaire. The new questionnaire was designed as an automated instrument. In contrast, the old Current Population Survey questionnaire used for the Current Population Survey CATI Phase-in Project was designed as a paper instrument. Field interviewers were required to memorize complicated skip patterns. To minimize respondent burden, both versions of the Current Population Survey questionnaire are designed for an average interview length of twenty minutes. Using an automated questionnaire, an interviewer can collect more (and more detailed) information in the same amount of time, since she no longer has to determine the path of the interview. Besides the automation difference, the wording of the labor force questions differs between the two questionnaires.

Parallel Survey interviews were conducted using the same questionnaire both in the field interviews (using a laptop computer) and in the CATI facilities. In contrast, the Current Population Survey CATI Phase-in interviews used two different versions of the old questionnaire: a paper version for the field interviews; and an automated version, with a slightly modified lead-in labor force question for the CATI interviews.

Given these questionnaire differences, and the caveats about the Parallel Survey split panel data, we view our results as preliminary. Instead, we recommend pursuing this examination using one and two-sample McNemar techniques on the new Current Population Survey split panel data, which uses the old CATI Phase-in design and the redesigned, fully automated questionnaire.

## 4. CONCLUSION

We have presented two modifications of the one and two-sample McNemar tests using complex survey data, with applications from the unlinked data modification. If the survey does not have a longitudinal design, then the application using the linked data will have an unknown variance/covariance structure and will include a variance component due to matching error. In this case, using the unlinked data makes sense with respect to the model's interpretation, although the statistic based on the (unlinked) estimates of marginal probabilities may be inferior to a well-developed linked model. If the survey has a longitudinal design, then the first method may be preferred, as it is a straight-forward extension of the traditional test, and consequently, the interpretation is equivalent to the textbook interpretation.

The two-sample McNemar test is not the sole approach one might use in the situation described in section 2.2.2. Another approach to the unlinked form of this problem would be to use a log-linear model for a  $2 \times 2 \times 2$  contingency table as in Rao and Scott (1984). In either case, there are trade-offs. The interpretation of the McNemar test is intuitive: it is a cause and effect model, or a repeated measures type of experimental design. The  $2 \times 2 \times 2$  contingency table model's interpretation is perhaps less intuitive. Note, however, that the test statistic for the McNemar tests are "Wald-like" statistics, which are often considered to be less efficient than the chi-squared type, *e.g.*, Fay (1985). It is also worth noting that unlike the Rao-Scott formulation, the approach described in this paper makes explicit provisions for the use of linked data.

Areas for future research include investigations into the power of these tests in the context of complex sample data, variance/covariance estimation for linked data including matching error variance contributions, and the difference in efficiency in the two approaches. In data analytical applications, one and two-sample McNemar tests seem to have uses in comparing aspects of different survey methods or effects on responses within a method over time. The approach is nonparametric in its conception; when the approximation is good, it avoids pitfalls that may be associated with model-based tests.

## ACKNOWLEDGEMENTS

We thank James Hartman, Alfredo Navarro, James Roebuck, Lynn Weidman, the referee, and the editors for their useful comments. We also thank Sue Chandler for her typing

of this paper. This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.

## APPENDIX

For the unlinked data modification of the McNemar Test,  $(p_2 - p_1)$  is estimated by  $X_{+,1}/N_1 + X_{+,2}/N_2$  where  $X_{+,1}$ ,  $X_{+,2}$ ,  $N_1$ , and  $N_2$  are weighted estimates, and

$$\begin{aligned} \widehat{\text{Var}}(p_2 - p_1) &= \left[ \frac{X_{+,1}}{N_1} \right]^2 \left[ \frac{\text{Var}(X_{+,1})}{X_{+,1}^2} - \frac{\text{Var}(N_1)}{N_1^2} \right] \\ &+ \left[ \frac{X_{+,2}}{N_2} \right]^2 \left[ \frac{\text{Var}(X_{+,2})}{X_{+,2}^2} - \frac{\text{Var}(N_2)}{N_2^2} \right] \\ &- \frac{2X_{+,1}X_{+,2}}{N_1N_2} \left[ \frac{\text{Cov}(X_{+,1}, X_{+,2})}{X_{+,1}X_{+,2}} - \frac{\text{Var}(N_1)}{N_1^2} \right. \\ &\quad \left. - \frac{\text{Var}(N_2)}{N_2^2} + \frac{\text{se}(N_1)\text{se}(N_2)}{N_1N_2} \right]. \end{aligned}$$

In this appendix we discuss the derivation of the covariance term in the variance estimate, considering only the unlinked data.

Consider the within-panel correlation

$$\text{Cov}(X_{+,1}, X_{+,2}) = \text{Cov} \left( \sum_{j=1,5} X_{1,j}, \sum_{j=2,6} X_{2,j} \right) \quad (\text{A1})$$

where  $X_{i,j}$  is a weighted sample level for month  $i$ , month-in-sample (MIS)  $j$ . Note that  $X_{1,j}$  and  $X_{2,j+1}$  are from the same rotation group unless  $j = 4$  since a rotation group is out of sample for eight months after being in for four.

We assumed that the correlations between  $X_{i,j}$  and  $X_{k,m}$  can be decomposed into three separate categories:

1) A within-rotation-group correlation,

$$\text{Cov}(X_{i,j}, X_{i+1,j+1}) = r_1, \text{ when } j = 1, 2, 3, 5, 6, 7.$$

2) A within-month-between-rotation group correlation,

$$\text{Cov}(X_{i,j}, X_{i,k}) = \omega, \text{ } k \neq j, \text{ and}$$

3) A between-rotation-group between-month correlation.

$$\text{Cov}(X_{i,j}, X_{i+1,k}) = \gamma, \text{ } k \neq j+1 \text{ or } j = 3.$$

Replicate estimates of these correlations were available.

The covariance in (A1) becomes

$$\begin{aligned} \text{Cov}(X_{+,1}, X_{+,2}) &= \text{Cov}(X_{1,1} + X_{1,5}, X_{2,2} + X_{2,6}) \\ &= \text{Cov}(X_{1,1}, X_{2,2}) + \text{Cov}(X_{1,1}, X_{2,6}) + \\ &\quad \text{Cov}(X_{1,5}, X_{2,2}) + \text{Cov}(X_{1,5}, X_{2,6}) \\ &= 2(r_1 + \gamma) \text{Var}(X_{i,j}), \end{aligned} \quad (\text{A2})$$

using the simplifying assumption that  $\text{Var}(X_{i,j})$  is constant for all  $i$  and  $j$ . The variance for a full month's estimate,  $\text{Var}(\sum_{j=1}^8 X_{i,j})$  is available in the form of a generalized variance function (GVF). We use this estimate to calculate  $\text{Var}(X_{i,j})$  by applying the following derivation:

$$\begin{aligned} \text{Var} \left( \sum_{j=1}^8 X_{i,j} \right) &= \sum_j \sum_k \text{Cov}(X_{i,j}, X_{i,k}) \\ &= \sum_j \text{Var}(X_{i,j}) + \sum_{j \neq k} \text{Cov}(X_{i,j}, X_{i,k}) \\ &= (8 + 56\omega) \text{Var}(X_{i,j}) \end{aligned}$$

so

$$\text{Var}(X_{i,j}) = (8 + 56\omega)^{-1} \text{Var} \left( \sum_{j=1}^8 X_{i,j} \right). \quad (\text{A3})$$

## REFERENCES

- ADAMS, D.E. (1991). Current population survey month-in-sample (MIS) bias index research. Internal memorandum, Demographic Statistical Methods Division, U.S. Bureau of the Census, Washington, DC.
- DONNER, A., and LI, K.Y.R. (1990). The relationship between chi-square statistics from matched and unmatched analyses. *Journal of Clinical Epidemiology*, 43, 827-831.
- FAY, R. (1990). VPLX: Variance estimates for complex samples. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.
- FAY, R. (1985). A jackknifed chi-squared test for complex samples. *Journal of the American Statistical Association*, 80, 148-157.
- FEUER, E.J., and KESSLER, L.G. (1989). Test statistic and sample size for a two-sample McNemar test. *Biometrics*, 45, 629-636.
- FISHER, R., ROBISON, E., THOMPSON, J., and WELCH, M. (1993). Variance estimation in the current population survey overlap test. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.
- FISHER, R., and McGUINNESS, R. (1993). Correlations and adjustment factors for current population survey. Internal memorandum, Demographic Statistical Methods Division, U.S. Bureau of the Census, Washington, DC.
- HOGUE, C. (1984). History of the problems encountered estimating gross flows. *Proceedings of the Conference on Gross Flows in Labor Force Statistics*, 1-7.
- JABINE, T.B., and SCHEUREN, F.J. (1986). Record linkages for statistical purposes: Methodological Issues. *Journal of Official Statistics*, 2, 3, 255-277.



- McNEMAR, Q. (1947). Note on the sampling error of the differences between correlated proportions of percentages. *Psychometrika*, 12, 153-157.
- MARASCUILO, L.A., OMELICH, C.L., and GOKHALE, D.V. (1988). Planned and post hoc methods for multiple-sample McNemar (1947) tests with missing data. *Psychological Bulletin*, 103, 238-245.
- RAO, J.N.K., and WU, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231-241.
- RAO, J.N.K., and SCOTT, A.J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *The Annals of Statistics*, 12, 46-60.
- SHOEMAKER, H.H. (1993). Results from the current population survey CATI phase-in project. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.
- STASNY, E.A., and FIENBERG, S.E. (1984). Some stochastic models for estimating gross flows in the presence of nonrandom nonresponse. *Proceedings of the Conference on Gross Flows in Labor Force Statistics*, 25-39.
- THOMPSON, J. (1994). Mode effects analysis of labor force estimates. Current Population Survey Overlap Analysis Team Technical Report 3, U.S. Bureau of the Census, Washington, DC.
- THOMPSON, J., and FISHER, R. (1994). Two sample McNemar test for complex surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.





# Stability Measures for Variance Component Estimators Under a Stratified Multistage Design

J.L. ELTINGE and D.S. JANG<sup>1</sup>

## ABSTRACT

In work with sample surveys, we often use estimators of the variance components associated with sampling within and between primary sample units. For these applications, it can be important to have some indication of whether the variance component estimators are stable, *i.e.*, have relatively low variance. This paper discusses several data-based measures of the stability of design-based variance component estimators and related quantities. The development emphasizes methods that can be applied to surveys with moderate or large numbers of strata and small numbers of primary sample units per stratum. We direct principal attention toward the design variance of a within-PSU variance estimator, and two related degrees-of-freedom terms. A simulation-based method allows one to assess whether an observed stability measure is consistent with standard assumptions regarding variance estimator stability. We also develop two sets of stability measures for design-based estimators of between-PSU variance components and the ratio of the overall variance to the within-PSU variance. The proposed methods are applied to interview and examination data from the U.S. Third National Health and Nutrition Examination Survey (NHANES III). These results indicate that the true stability properties may vary substantially across variables. In addition, for some variables, within-PSU variance estimators appear to be considerably less stable than one would anticipate from a simple count of secondary units within each stratum.

**KEY WORDS:** Between-PSU variance; Complex sample design; Degrees of freedom; Diagnostic; Design-based analysis; Satterthwaite approximation; Stratum collapse; U.S. Third National Health and Nutrition Examination Survey (NHANES III); Within-PSU variance.

## 1. INTRODUCTION

In work with sample surveys, it is often desirable to have good estimates of the variance components attributable to sampling within and between primary sample units (PSUs). For example, the magnitude of an estimated within-PSU variance, relative to a between-PSU variance, may influence decisions regarding sample allocation and related design issues (*e.g.*, Hansen *et al.* 1953, Chapter 7). Similar relative-magnitude properties affect the bias of certain variance estimators derived under simplifying assumptions regarding the sample design (*e.g.*, Korn and Graubard 1995, p. 278-279, 287; and Wolter 1985, p. 44-46). Also, some survey analysts have a general interest in identification of surveys and variables for which the between-PSU component of variance is substantially greater than zero. See, *e.g.*, Herzog and Scheuren (1976, p. 398) and Wolter (1985, p. 47) for related comments. In addition, Jang and Eltinge (1996) give an example for which there is some interest in the within-PSU variances by themselves.

In some application work, estimates of within-PSU variances and related quantities are reported with the apparent assumption that the estimates are stable, *i.e.*, have relatively low variances. This paper shows that it can be important to carry out data-based checks of this assumption of stability, and that some relatively simple checking methods follow from standard design-based ideas. We emphasize methods that can be applied to designs with a moderate or large number of strata and a small number of PSUs selected per stratum.

Subsection 2.1 reviews the relevant estimators of within-PSU variances and overall stratum-level variances. Subsection 2.2 identifies two distinct components of the variance of the within-PSU variance estimator. Subsection 2.3 presents simple design-based estimators of the variances of two within-PSU variance estimators. Section 3 develops two related degrees-of-freedom measures.

Section 4 examines the extent to which related design-based methods can be used to assess the stability of quantities that depend both on the within-PSU variance estimator and on the overall stratum-level variance estimator. Principal attention is directed toward an estimator of the between-PSU variance and an estimator of the ratio of the overall stratum-level variance divided by the within-PSU variance. Section 4.2 discusses one set of methods based on the stability measures from Section 2 and some moderately restrictive moment assumptions. Section 4.3 outlines a second set of methods based on stratum collapse.

Section 5 applies the main ideas of Sections 2 through 4 to variance estimates computed for the U.S. Third National Health and Nutrition Examination Survey. Section 5 also uses a simple simulation-based method to assess the consistency of the observed measures with standard assumptions regarding variance estimator stability. The Section 5 results suggest that the true stability of within-PSU variance estimators can be substantially less than anticipated from a simple count of the number of secondary units contributing to each PSU. In addition, the results indicate that the stability properties of

<sup>1</sup> J.L. Eltinge and D.S. Jang, Department of Statistics, Texas A&M University, College Station, TX 77843-3143, U.S.A.

within-PSU variance estimators and related quantities can vary substantially across different variables collected in the same survey. Section 6 gives additional comments on the methods and empirical results presented here.

## 2. WITHIN-PSU AND OVERALL STRATUM-LEVEL VARIANCE ESTIMATORS

### 2.1 General Notation

In principle, we could use either design-based or model-based methods to examine within-PSU and between-PSU variance components. The present work will take a design-based approach. This is consistent with some related previous literature, *e.g.*, Wolter (1985, p. 40-41, 47). The design-based approach will be especially useful in highlighting some strengths and limitations of the proposed stability-assessment methods. For example, in Section 2.3 this approach will give us some indication of specific design features that may affect variance estimator stability. Also, in Section 4 the design-based approach will help to clarify the extent to which certain moment restrictions are needed to justify one set of stability measures.

Following the notation and ideas in Wolter (1985, p. 43-47), consider a stratified multistage sample design with  $L$  strata and with  $N_h$  primary sampling units (PSUs) contained in stratum  $h = 1, 2, \dots, L$ . We select  $n_h$  PSUs with replacement and with per-draw selection probabilities  $p_{hi}$ . Within selected PSU  $(h, i)$ , we select  $n_{hi}$  secondary sample units (SSUs) with replacement and with per-draw selection probabilities  $p_{hij}$ . Further subsampling is carried out within a selected SSU to obtain  $n_{hij}$  individual elements for interview or examination. The stability-assessment methods developed here are intended primarily for designs with moderate or large  $L$ , relatively small  $n_h$  (*e.g.*,  $n_h = 2$ ), and relatively large  $n_{hi}$ . Designs with these characteristics are often used in large household interview surveys, *e.g.*, the health survey discussed in Section 4.

We will focus on estimation of a population total  $Y = \sum_{h=1}^L Y_h$ , where  $Y_h = \sum_{i=1}^{N_h} Y_{hi}$ ,  $Y_{hi} = \sum_{j=1}^{N_{hi}} \sum_{k=1}^{N_{hij}} Y_{hijk}$ ,  $Y_{hijk}$  is a survey item for element  $k$  in SSU  $j$  in PSU  $i$  in stratum  $h$ ,  $N_{hi}$  is the number of SSUs in PSU  $(h, i)$ , and  $N_{hij}$  is the number of elements in SSU  $(h, i, j)$ . Extensions to nonlinear functions of population totals are straightforward and will be considered in the applications in Section 5. A standard design-based estimator of  $Y$  is  $\hat{Y} = \sum_{h=1}^L \hat{Y}_h$  where

$$\hat{Y}_h = \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} \sum_{k=1}^{n_{hij}} w_{hijk} y_{hijk}, \quad (2.1)$$

$w_{hijk}$  is the customary weight derived from selection probabilities and sample sizes to ensure unbiased estimation of each  $Y_h$ , and the lower-case terms  $y_{hijk}$  denote sample observations. In subsequent work, it will be useful to rewrite expression (2.1) as

$$\hat{Y}_h = n_h^{-1} \sum_{i=1}^{n_h} p_{hi}^{-1} \hat{Y}_{hi},$$

where  $\hat{Y}_{hi} = n_{hi}^{-1} \sum_{j=1}^{n_{hi}} z_{hij}$  and  $z_{hij} = n_h n_{hi} p_{hi} \sum_{k=1}^{n_{hij}} w_{hijk} y_{hijk}$ .

### 2.2 Within- and Between-PSU Variances

Throughout this discussion, expectations and variances will be defined with respect to the sample design. Under the conditions stated above, the variance of  $\hat{Y}$  is  $V(\hat{Y}) = \sum_{h=1}^L V_h$ , where  $V_h = V_{Bh} + V_{Wh}$ ,  $V_{Bh} = V(n_h^{-1} \sum_{i=1}^{n_h} p_{hi}^{-1} Y_{hi})$ ,  $V_{Wh} = n_h^{-1} \sum_{i=1}^{n_h} p_{hi}^{-1} \sigma_{2hi}^2$ , and  $\sigma_{2hi}^2 = V(\hat{Y}_{hi} - Y_{hi} | h, i)$ ; see, *e.g.*, Wolter (1985, p. 42). Note especially that  $Y_{hi}$  is the true population total for selected PSU  $(h, i)$ , and that  $\sigma_{2hi}^2$  reflects the variability in  $\hat{Y}_{hi} - Y_{hi}$  attributable to subsampling at the SSU and finer levels.

A customary unbiased estimator of the overall stratum-level variance  $V_h$  is

$$\hat{V}(\hat{Y}_h) = n_h^{-1} (n_h - 1)^{-1} \sum_{i=1}^{n_h} (p_{hi}^{-1} \hat{Y}_{hi} - \hat{Y}_h)^2,$$

and the corresponding estimator of  $V(\hat{Y}) = \sum_{h=1}^L V(\hat{Y}_h)$  is  $\hat{V}(\hat{Y}) = \sum_{h=1}^L \hat{V}(\hat{Y}_h)$ .

Now consider estimation of the within-PSU variance  $V_{Wh}$ . Since  $\hat{Y}_{hi}$  is a sample mean of the independent and identically distributed terms  $z_{hij}$ , standard arguments show that for a given PSU  $(h, i)$ , an unbiased estimator of  $\sigma_{2hi}^2$  is  $\hat{\sigma}_{2hi}^2 = n_{hi}^{-1} (n_{hi} - 1)^{-1} \sum_{j=1}^{n_{hi}} (z_{hij} - \hat{Y}_{hi})^2$ . Thus, an unbiased estimator of  $V_{Wh}$  is

$$\hat{V}_{Wh} = \sum_{i=1}^{n_h} n_h^{-2} p_{hi}^{-2} \hat{\sigma}_{2hi}^2 = \sum_{i=1}^{n_h} n_{hi}^{-1} (n_{hi} - 1)^{-1} \sum_{j=1}^{n_{hi}} (x_{hij} - \bar{x}_{hi})^2,$$

where  $x_{hij} = n_{hi} \sum_{k=1}^{n_{hij}} w_{hijk} y_{hijk}$  and  $\bar{x}_{hi} = n_{hi}^{-1} \sum_{j=1}^{n_{hi}} x_{hij}$ . Note that the latter expression for  $\hat{V}_{Wh}$  uses only sample sizes, the observations  $y_{hijk}$  and the customary weights  $w_{hijk}$ .

### 2.3 The Variance of $\hat{V}_{Wh}$

A direct modification of standard conditional-moment arguments shows that the variance of  $\hat{V}_{Wh}$  is  $\gamma_{Bh} + \gamma_{Wh}$ , where

$$\gamma_{Bh} = V(n_h^{-2} \sum_{i=1}^{n_h} p_{hi}^{-2} \sigma_{2hi}^2)$$

and

$$\gamma_{Wh} = n_h^{-3} \sum_{i=1}^{n_h} p_{hi}^{-3} V(\hat{\sigma}_{2hi}^2 | h, i).$$

Thus, the variance of  $\hat{V}_{Wh}$  itself depends on a sum of between- and within-PSU variances, and the relative magnitudes of  $\gamma_{Bh}$  and  $\gamma_{Wh}$  depend on trade-offs among  $\sigma_{2hi}^2$ ,  $p_{hi}$  and  $n_{hi}$ . For example, under regularity conditions, the terms  $V(\hat{\sigma}_{2hi}^2 | h, i)$  are approximately inversely proportional to  $n_{hi}$ . Thus, if the  $n_{hi}$  are uniformly large within stratum  $h$ , then  $\gamma_{Wh}$



may be relatively small. Also, if the terms  $p_{hi}^{-2} \sigma_{2hi}^2$  are approximately constant within a given stratum, then  $\gamma_{Bh}$  may be relatively small. Conversely, marked heterogeneity of  $p_{hi}^{-2} \sigma_{2hi}^2$  may inflate  $\gamma_{Bh}$  and thus inflate  $V(\hat{V}_{wh})$  as well.

In addition, note that under the stated design conditions,  $\hat{V}_{wh}$  is the sample mean of the independent and identically distributed terms  $n_h^{-1} p_{hi}^{-2} \delta_{2hi}^2$ . Thus, an unbiased estimator of the variance of  $\hat{V}_{wh}$  is

$$\tilde{V}(\hat{V}_{wh}) = n_h^{-1} (n_h - 1)^{-1} \sum_{i=1}^{n_h} (n_h^{-1} p_{hi}^{-2} \delta_{2hi}^2 - \hat{V}_{wh})^2. \quad (2.2)$$

Some applications focus on the full-population level, rather than on individual strata, and so the “within-PSU” contribution of interest is the sum of the within-PSU variances,  $V_w = \sum_{h=1}^L V_{wh}$ . Under the conditions given above, an unbiased estimator of  $V_w$  is  $\hat{V}_w = \sum_{h=1}^L \hat{V}_{wh}$ . Also, since our sampling and subsampling are independent across strata, we have  $V(\hat{V}_w) = \sum_{h=1}^L (\gamma_{Bh} + \gamma_{wh})$ , and an unbiased estimator of  $V(\hat{V}_w)$  is

$$\tilde{V}(\hat{V}_w) = \sum_{h=1}^L \tilde{V}(\hat{V}_{wh}).$$

Finally, note that the preceding development used the assumption of sampling with replacement at both the primary- and secondary-unit levels. Two applications of result (2.4.16) in Wolter (1985, p. 46) show that under mild conditions that hold for many, but not all, without-replacement designs,  $\hat{V}_{wh}$  will be unbiased or conservative for the true within-PSU variance; and  $\tilde{V}(\hat{V}_{wh})$  will be unbiased or conservative for the true variance of  $\hat{V}_{wh}$ . A formal technical statement and proof of this result is available from the authors.

## 2.4 Balanced Interpretation of Stability Measures

The remainder of this paper uses  $\tilde{V}(\hat{V}_{wh})$  and related quantities to assess the stability of variance-component estimators. In working with these results, it is useful to remember that data-based measures of variance estimator stability are justifiably viewed with some caution, because they are functions of fourth sample moments, and thus are themselves subject to a considerable amount of sampling variability. See, e.g., Fuller (1984, p. 111). This caution carries over to the proposed estimator  $\tilde{V}(\hat{V}_{wh})$  and to the related statistics discussed in Sections 3 and 4 below.

However, one should not overstate this caution to the point of making no attempt at data-based assessment of variance estimator stability. The estimator  $\tilde{V}(\hat{V}_{wh})$ , and the related measures in Sections 3 and 4, are relatively simple to compute, and provide diagnostics that can help to identify variables for which:

- the instability of  $\hat{V}_{wh}$  is especially problematic; or
- the instability of  $\hat{V}_{wh}$  has a substantial effect on the precision of estimators of the relative magnitudes of between-PSU and within-PSU variances.

Consequently, interpretation of specific values of  $\tilde{V}(\hat{V}_{wh})$  and related stability measures should reflect a balance between the abovementioned general caution and a recognition of their potential diagnostic value.

## 3. TWO STABILITY MEASURES FOR WITHIN-PSU VARIANCE ESTIMATORS

### 3.1 Degrees-of-Freedom Diagnostics for Variance Estimator Stability

Some analysts prefer to express variance estimator stability through “degrees of freedom” measures related to the Satterthwaite (1941, 1946) approximation. To introduce this idea, consider a general variance estimator  $\hat{V}$ , and note that  $\{E(\hat{V})\}^{-1} d\hat{V}$  has the same first and second moments as a chi-square random variable on  $d$  degrees of freedom, where  $d$  is the solution to the equation,

$$2\{E(\hat{V})\}^2 - V(\hat{V})d = 0.$$

If the distribution of  $\{E(\hat{V})\}^{-1} d\hat{V}$  is indeed well approximated by a chi-square distribution, then  $d$  may be viewed fairly literally as a “degrees of freedom” term. Otherwise,  $d$  can be viewed as twice the inverse of the squared coefficient of variation of  $\hat{V}$ . In either case,  $d$  has a certain appeal because it is scale-free, and can be tied fairly directly to notions of “effective sample size” in the evaluation of variance estimator performance. Subsection 3.3 gives related comments for two special cases.

Given an unbiased estimator  $\tilde{V}(\hat{V})$  of the variance of  $\hat{V}$ , one may compute a “degrees of freedom” estimator  $\hat{d}$  as the solution to the unbiased estimating equation

$$2\{\hat{V}^2 - \tilde{V}(\hat{V})\} - \tilde{V}(\hat{V})d = 0, \quad (3.1)$$

i.e.,  $\hat{d} = \{\tilde{V}(\hat{V})\}^{-1} 2\hat{V}^2 - 2$ . Under mild regularity conditions,  $\hat{d}^{-1} \hat{d}$  converges in probability to one, provided  $\{V(\hat{V})\}^{-1} \tilde{V}(\hat{V})$  and  $\{E(\hat{V})\}^{-1} \hat{V}$  both converge in probability to one.

### 3.2 Degrees-of-Freedom Diagnostics for Pooled and Stratum-Level Estimators of Within-PSU Variances

We can apply these general degrees-of-freedom ideas to the within-PSU variance estimators  $\hat{V}_{wh}$  and  $\hat{V}_w$  developed in Section 2. First consider the case in which there is intrinsic interest in the stability of individual stratum-level estimators  $\hat{V}_{wh}$ . The associated “degrees of freedom” measure is  $d_{wh} = \{V(\hat{V}_{wh})\}^{-1} 2V_{wh}^2$ . For designs with large  $n_h$ , one may use (3.1) to compute estimators  $\hat{d}_{wh} = \{\tilde{V}(\hat{V}_{wh})\}^{-1} 2\hat{V}_{wh}^2 - 2$  separately for each stratum. For designs with small  $n_h$  (e.g.,  $n_h = 2$  for each stratum), the estimator  $\hat{d}_{wh}$  itself may be very unstable.

Consequently, it also is useful to consider the alternative combined estimator

$$\hat{d}_{w0} = \left\{ \sum_{h=1}^L \tilde{V}(\hat{V}_{wh}) \right\}^{-1} 2 \sum_{h=1}^L \hat{V}_{wh}^2 - 2,$$

under the assumption that all  $d_{wh}$  equal a common value  $d_{w0}$ .

Now consider the pooled within-PSU variance estimator  $\hat{V}_w$  developed in Section 2.3. The resulting “degrees of freedom” measure is  $d_{wF} = \left\{ \sum_{h=1}^L V(\hat{V}_{wh}) \right\}^{-1} 2V_w^2$ , and expression (3.1) suggests the estimator

$$\hat{d}_{wF} = \left\{ \sum_{h=1}^L \tilde{V}(\hat{V}_{wh}) \right\}^{-1} 2\hat{V}_w^2 - 2.$$

### 3.3 Comparison of $d_{w0}$ and $d_{wF}$ to Direct SSU Counts

To interpret  $\hat{d}_{w0}$  and  $\hat{d}_{wF}$  as stability measures, consider the following idealized setting. Assume that for all  $h$ , the PSU counts  $n_h$  are equal to a common value  $n_1$ , say; and that for all  $h$  and  $i$ , the SSU counts  $n_{hi}$  are equal to a common value  $n_{11}$ . In addition, assume that the terms  $p_{hi}^{-2} \sigma_{2hi}^2$  are constant within each stratum; and that, conditional on  $(h, i)$ , each  $\sigma_{2hi}^{-2} (n_{11} - 1) \delta_{2hi}^2$  is distributed as a chi-square random variable on  $n_{11} - 1$  degrees of freedom. Then routine arguments show that  $d_{w0} = n_1(n_{11} - 1)$ . If the preceding assumptions are satisfied approximately, and if the product  $n_1(n_{11} - 1)$  is large (greater than 40, say), then a data user may be inclined to view  $\hat{V}_{wh}$  as relatively stable, or equivalently, to view the errors  $\hat{V}_{wh} - V_{wh}$  as negligible. This appears to be the reasoning used implicitly when estimates  $\hat{V}_{wh}$  are treated as known values in design or analysis work. However, the application in Section 5 will give some examples for which this reasoning is problematic, so that evaluation of the estimates  $\hat{d}_{w0}$  is important.

Also, under the idealized conditions described above, and under the additional assumption that the  $V_{wh}$  are all equal, we have  $d_{wF} = Ln_1(n_{11} - 1)$ .

## 4. COMPARISON OF WITHIN-PSU AND OVERALL STRATUM-LEVEL VARIANCES

### 4.1 Estimators of Between-PSU Variances and Related Variance Ratios

Section 1 cited some applications that hinge on the magnitude of  $V_{wh}$  relative to  $V_h$ . The specifics of the relative-magnitude comparisons vary with the individual application, but interest generally focuses on differences or ratios. For example, recall that  $V_{Bh} = V_h - V_{wh}$  and define the overall between-PSU variance term  $V_B = \sum_{h=1}^L V_{Bh}$ . In addition, note that unbiased estimators of  $V_{Bh}$  and  $V_B$  are  $\hat{V}_{Bh} = \hat{V}_h - \hat{V}_{wh}$  and  $\hat{V}_B = \sum_{h=1}^L \hat{V}_{Bh}$  respectively.

Similarly, define the ratio  $R_{wV} = V_w^{-1} V(\hat{Y})$ , the magnitude of the overall variance  $V(\hat{Y})$  relative to the within-PSU contribution  $V_w$ . A direct estimator of  $R_{wV}$  is  $\hat{R}_{wV} = \hat{V}_w^{-1} \hat{V}(\hat{Y})$ .

Note that if  $V_{wh}^{-1} V_h = R_{wV}$  for all  $h$ , then  $\hat{R}_{wV}$  could also be viewed as a pooled estimator of this common stratum-level ratio.

For both  $\hat{V}_B$  and  $\hat{R}_{wV}$ , stability assessment involves the variance of  $\hat{V}_h$  and the covariance of  $\hat{V}_{wh}$  with  $\hat{V}_h$ . Estimation of these moments can be somewhat problematic for surveys that select small numbers of PSUs from each stratum. We consider two approaches to resolving this problem. Section 4.2 uses moderate restrictions on the moment structure of  $(\hat{V}_{wh}, \hat{V}_h)$  to develop estimators  $V(\hat{V}_h)$  and related quantities. Section 4.3 uses stratum collapse to develop alternative stability measures.

### 4.2 Stability Measures Based on $\tilde{V}(\hat{V}_{wh})$ and Moment Conditions

#### 4.2.1 Moment Conditions

Under moderate moment restrictions, we can estimate the variance of  $\hat{V}_h$  directly from  $\hat{V}_h$  itself. Specifically, assume that the variance of  $\hat{V}_h$  equals  $(n_h - 1)^{-1} 2V_h^2$ ; this would hold, e.g., under the standard assumption that  $V_h^{-1} (n_h - 1) \hat{V}_h$  is distributed as a chi-square random variable on  $n_h - 1$  degrees of freedom. As in Sections 2 and 3, we continue to assume that  $\hat{V}_h$  is unbiased for  $V_h$ . Then routine moment arguments show that  $(n_h + 1)^{-1} 2\hat{V}_h^2$  is an unbiased estimator of the variance of  $\hat{V}_h$ .

In the remainder of Section 4.2, we will also assume that  $\text{Cov}(\hat{V}_{wh}, \hat{V}_h) = 0$ . Routine conditional-moment arguments show that this will hold if the terms  $p_{hi}^{-2} \sigma_{2hi}^2$  are equal within a given stratum; and if, conditional on  $(h, i, j)$ , the SSU-level estimates  $x_{hij}$  are normally distributed, so that  $\delta_{2hi}^2$  is conditionally independent of  $\hat{V}_{hi}$ .

#### 4.2.2 Stability Measures

Under the conditions stated in Section 4.2.1, unbiased estimators of  $V(\hat{V}_{Bh})$  and  $V(\hat{V}_B)$  are

$$\tilde{V}(\hat{V}_{Bh}) = (n_h + 1)^{-1} 2\hat{V}_h^2 + \tilde{V}(\hat{V}_{wh}) \quad (4.1)$$

and  $\tilde{V}(\hat{V}_B) = \sum_{h=1}^L \tilde{V}(\hat{V}_{Bh})$ , where  $\tilde{V}(\hat{V}_{wh})$  is defined in expression (2.2). Also, under the same conditions routine ratio-estimation arguments lead to the variance estimator

$$\tilde{V}(\hat{R}_{wV}) = \hat{V}_w^2 \sum_{h=1}^L \left\{ (n_h + 1)^{-1} 2\hat{V}_h^2 + \hat{R}_{wV}^2 \tilde{V}(\hat{V}_{wh}) \right\}. \quad (4.2)$$

### 4.3 Alternative Stability Measures Based on Stratum Collapse

The assumptions of Section 4.2.1 may be problematic in some applications. For example, for some survey designs and variables, the SSU-level estimators  $x_{hij}$  may have markedly nonnormal distributions, so the assumption  $\text{Cov}(\hat{V}_{wh}, \hat{V}_h) = 0$  may not hold. For these cases, one may consider the use of stratum collapse to produce alternative estimators of  $V(\hat{V}_B)$  and  $V(\hat{R}_{wV})$ .



Specifically, partition the set of  $L$  strata into  $G$  prespecified groups, with  $L_g$  strata contained in group  $S_g$ ,  $g = 1, \dots, G$ . With this new notation, note that

$$(\hat{V}(\hat{Y}), \hat{V}_W, \hat{V}_B) = \sum_{g=1}^G \sum_{h \in S_g} (\hat{V}_h, \hat{V}_{wh}, \hat{V}_{Bh}).$$

Standard stratum-collapse methods (e.g., Wolter 1985, Section 2.5) then lead to the alternative variance estimator,

$$V_{cs}^*(\hat{V}_B) = \sum_{g=1}^G (L_g - 1)^{-1} L_g \sum_{h \in S_g} D_{gh}^2,$$

where  $D_{gh} = \hat{V}_{Bh} - L_g^{-1} \sum_{j \in S_g} \hat{V}_{Bj}$ . Similarly, a collapsed-stratum variance estimator for  $\hat{R}_{WV}$  is,

$$V_{cs}^*(\hat{R}_{WV}) = (\hat{V}_W)^{-2} \sum_{g=1}^G (L_g - 1)^{-1} L_g \sum_{h \in S_g} C_{gh}^2,$$

where  $C_{gh} = (\hat{V}_h - \hat{R}_{WV} \hat{V}_{wh}) - L_g^{-1} \sum_{j \in S_g} (\hat{V}_j - \hat{R}_{WV} \hat{V}_{wj})$ .

In general, collapsed-stratum variance estimators require some care in interpretation; see, e.g., Rust and Kalton (1985), Wolter (1985, Section 2.5) and references cited therein. For example, collapsed-stratum variance estimators generally will be conservative. In addition, for cases with moderate  $L$ , the variance estimators  $V_{cs}^*(\hat{V}_B)$  and  $V_{cs}^*(\hat{R}_{WV})$  may themselves have limited stability.

## 5. APPLICATION TO THE U.S. THIRD NATIONAL HEALTH AND NUTRITION EXAMINATION SURVEY

### 5.1 Sample Design and Estimation Methods

The methods proposed in Sections 2 through 4 were applied to data from Phase I of the Third National Health and Nutrition Examination Survey (NHANES III). National Center for Health Statistics (1996) gives a general description of this survey, including special characteristics associated with Phase I (data collected between 1988 and 1991). For the present discussion, three aspects are of special interest. First, variance estimators were constructed on the basis of a collapsed design involving  $L = 22$  strata (large groups of counties), with two primary sample units (generally individual counties) selected per stratum. Second, each selected PSU had a relatively large number of selected SSUs (generally groups of city blocks, or similar rural areas). The number of selected SSUs within each stratum ranged from 30 to 63, with a mean of 45.8.

Third, additional subsampling within each SSU led to selection of the survey elements (individual noninstitutionalized U.S. civilians). Each selected person was asked to respond to a health questionnaire and to participate in a detailed medical examination. Twelve of the resulting variables are listed in Table 1.

Standard weighted ratio estimates  $\hat{\theta}$  were computed for the population means of each of the twelve variables listed in Table 1. The first two columns of Table 2 present the corresponding variance estimates  $\hat{V}(\hat{\theta})$  and  $\hat{V}_W$ . As part of a larger study of the within-PSU variances  $V_{wh}$  discussed in Jang and Eltinge (1996), there was considerable interest in the stability of the individual estimates  $\hat{V}_{wh}$ . Since we had  $n_h = 2$  for each stratum, the reasoning in Section 3.2 indicated that it was not feasible to examine the individual terms  $\hat{d}_{wh}$ . Consequently, Section 5.2 will examine the pooled measure  $\hat{d}_{w0}$  of the stability of the  $\hat{V}_{wh}$  and will also present some related simulation-based tests and diagnostic plots.

**Table 1**  
Twelve NHANES III Variables

Variable name	Description
HAE2	Told by health professional that you had hypertension (indicator variable)
HAE7	Told by health professional that your blood cholesterol was high (indicator variable)
HAD1	Told by health professional that you had diabetes (indicator variable)
HAR3	Do you smoke cigarettes now?
BMPHT	Height
BMPWT	Weight
HDRESULT	HDL cholesterol
TCRESULT	Serum total cholesterol
LEAD	Blood lead, in micrograms per deciliter
log(LEAD)	Natural logarithm of blood lead
BP1K1	Systolic blood pressure
BP1K5	Diastolic blood pressure

**Table 2**  
Variance Estimates and Stability Measures for  
Twelve NHANES III Variables

Variable name	$\hat{V}_W$	$\hat{V}(\hat{Y})$	$\hat{d}_{w0}$	$\hat{d}_{WF}$
HAE2	0.0000385	0.0000511	23.7	425.8
HAE7	0.0000821	0.000135	13.6	225.6
HAD1	0.00000956	0.00000749	8.8	160.6
HAR3	0.000122	0.000205	6.4	125.8
BMPHT	0.0223	0.0416	15.3	275.1
BMPWT	0.104	0.122	8.6	139.2
HDRESULT	0.0743	0.163	11.5	196.2
TCRESULT	0.590	0.860	21.2	353.9
LEAD	0.00388	0.00657	2.8	48.8
log(LEAD)	0.000211	0.000678	10.5	174.9
BP1K1	1.073	2.896	1.0	26.5
BP1K5	0.252	0.217	17.2	52.9

In addition, there was interest in the extent to which the variances of the  $\hat{V}_{wh}$  contributed to the variances of the pooled quantities  $\hat{V}_B$  and  $\hat{R}_{WV}$ . Section 5.3 explores this question.

## 5.2 Within-PSU Variance Estimates and Associated Stability Measures

### 5.2.1 Comparison Across Variables

The final two columns of Table 2 report the degrees-of-freedom estimates  $\hat{d}_{w0}$  and  $\hat{d}_{wF}$  for the twelve NHANES III variables. Note especially that the stratum-level stability measures  $\hat{d}_{w0}$  are relatively low, compared to the mean of 45.8 SSUs per stratum. For example, all of the variables have  $\hat{d}_{w0}$  less than 24, and five (HAD1, HAR3, BMPWT, LEAD and BP1K1) have  $\hat{d}_{w0}$  less than 10. Due to the interest in the  $\hat{d}_{w0}$  described above, this led to two general questions.

- (1) Are the observed  $\hat{d}_{w0}$  consistent with the nominal degrees-of-freedom value  $d_{w0}$  that one would anticipate from the direct SSU counts  $n_{h1} + n_{h2} - 2$ ?
- (2) Conversely, are the observed  $\hat{d}_{w0}$  consistent with distributional conditions that produce considerably smaller values of  $d_{w0}$ ?

Standard large-sample-theory-based tests for (1) and (2) would have depended on eighth sample moments, and thus were inadvisable in the present case, due to the relatively small values of  $L = 22$  and  $n_h = 2$ . Instead, the following simulation-based test was carried out.

### 5.2.2 Simulation-Based Interpretation of Stability Measures

This simulation work covers six cases involving different values of two terms. The first term, denoted  $d_{hi}$ , represents the degrees of freedom associated with the variance estimator  $\hat{\sigma}_{2hi}^2$  in PSU ( $h, i$ ). The second term, denoted  $R_{12}$ , is the ratio of the expressions  $p_{hi}^{-2} \sigma_{2hi}^2$  in the first and second sample PSUs in stratum  $h$ .

In each of the six cases discussed below, independent pseudorandom variables  $g_{hi}$  were generated from a chi-square distribution on  $d_{hi}$  degrees of freedom for  $h = 1, 2, \dots, 22$  and  $i = 1, 2$ . Re-scaled variables  $\hat{V}_{whi} = d_{hi}^{-1} V_{whi} g_{hi}$  were then computed, where  $V_{whi}$  is a random variable equal to one with probability one-half and equal to  $R_{12}$  with probability one-half. The random variables  $g_{hi}$  and  $V_{whi}$  are mutually independent. Finally, the sums  $\hat{V}_{wh} = \hat{V}_{wh1} + \hat{V}_{wh2}$  and the associated measures  $\hat{V}(\hat{V}_{wh})$ ,  $\hat{V}(\hat{V}_w)$  and  $\hat{d}_{w0}$  were computed. This was repeated 10,000 times.

Table 3 lists the values of  $d_{hi}$  and  $R_{12}$  covered in the six cases, and Table 4 lists the resulting simulated means,

standard deviations and quantiles for  $\hat{d}_{w0}$ . When interpreting the results for these cases, note that randomness of the  $g_{hi}$  corresponds to the estimation error in the  $\hat{\sigma}_{2hi}^2$  due to subsampling at the SSU and lower levels; and randomness of the  $V_{whi}$  reflects the variability of the  $p_{hi}^{-2} \sigma_{2hi}^2$  induced by sampling of PSUs within a given stratum.

**Table 3**  
Cases Covered for the Simulated Quantiles

Cases	$d$	$R_{12}$
1	22	1
2	Obs. Dist.	1
3	5	1
4	22	9
5	Obs. Dist.	9
6	5	9

Case 1 uses  $d_{hi} = 22$  and  $R_{12} = 1$ . Arguments from Section 3.3 show that the resulting  $\hat{V}_{wh}$  are distributed as constant multiples of a chi-square random variable with  $d_{w0} = 44$  degrees of freedom. Thus, for Case 1, the choice of  $d_{hi} = 22$  has led to simulated quantiles of  $\hat{d}_{w0}$  that are approximately those that one would anticipate from the mean SSU count of 45.8 observed for Phase I of NHANES III, under the setting described in Section 3.4. Note that even in this idealized Case 1, the relative variability of the  $\hat{d}_{w0}$  is fairly high.

Now compare the  $\hat{d}_{w0}$  reported in Table 2 to the simulated quantiles from Case 1. All twelve of the observed  $\hat{d}_{w0}$  fall below the 0.025 simulated quantile of 24.8; and ten of the twelve fall below the 0.005 quantile of 21.1. Thus, the  $\hat{d}_{w0}$  observed for the NHANES III variables are not consistent with a nominal  $d_{w0} = 44$  produced in the idealized setting covered by Case 1.

### 5.2.3 Simulation Under Alternative Conditions with Smaller $d_{w0}$

In general, the distribution of  $\hat{d}_{w0}$  may deviate from that observed under the idealized Case 1 due to: (a) variability in the true SSU counts  $n_{hi}$ ; (b) limited stability of the PSU-level estimates  $\hat{\sigma}_{2hi}^2$ ; and (c) heterogeneity of the true PSU-level terms  $\sigma_{2hi}^2$ . Cases 2 through 6 cover the combined effects of these three factors.

**Table 4**  
Simulated Quantiles for  $\hat{d}_{w0}$

Cases	Mean	S.D.	$q_{.005}$	$q_{.01}$	$q_{.025}$	$q_{.05}$	$q_{.10}$	$q_{.25}$	$q_{.50}$	$q_{.75}$	$q_{.90}$	$q_{.95}$	$q_{.975}$	$q_{.99}$	$q_{.995}$
1	48.9	17.7	21.1	22.5	24.8	27.4	30.7	36.7	45.5	57.4	71.2	81.5	92.6	108.5	122.1
2	48.3	17.5	20.7	21.9	24.2	26.8	29.9	36.3	45.2	56.6	70.2	80.3	92.0	106.2	118.0
3	11.3	4.7	4.1	4.5	5.1	5.6	6.4	8.0	10.3	13.5	17.3	20.0	23.0	26.8	30.1
4	5.5	2.7	1.4	1.6	2.0	2.3	2.7	3.7	5.0	6.8	8.9	10.5	12.1	14.8	16.7
5	5.5	2.7	1.4	1.6	1.9	2.3	2.7	3.7	5.0	6.7	8.9	10.6	12.1	14.1	16.1
6	3.5	2.1	0.7	0.8	1.0	1.2	1.5	2.1	3.0	4.4	6.0	7.4	8.8	11.2	12.6



The design for Case 2 was identical to that for Case 1, except that the  $d_{hi}$  were random variables, selected with equal probabilities and with replacement from the 44 values  $n_{hi} - 1$  corresponding to the 44 SSU counts  $n_{hi}$  in the original dataset. The resulting simulated quantiles of  $\hat{d}_{w0}$  are similar to those for Case 1.

Case 3 uses  $d_{hi} = 5$  and  $R_{12} = 1$ ; the resulting  $\hat{V}_{wh}$  are distributed as constant multiples of chi-square random variables with  $d_{w0} = 10$  degrees of freedom. The simulated quantiles for Case 3 were somewhat more consistent with the  $\hat{d}_{w0}$  observed for the NHANES III dataset. For example, ten of the twelve variables have  $\hat{d}_{w0}$  at or above the simulated 0.10 quantile of 6.4. However, two of the variables (lead and systolic blood pressure) had their  $\hat{d}_{w0}$  below the simulated 0.005 quantile for Case 3.

Cases 4 through 6 cover more extreme cases of instability, induced by use of the scale factor  $R_{12} = 9$ . A scale factor different from one introduces a component of variability associated with sampling of PSUs with unequal  $\sigma_{2hi}^2$ , and causes the  $\hat{V}_{wh}$  to have distributions outside of the rescaled chi-square family. Cases 4 through 6 use the same  $d_{hi}$  values used in Cases 1 through 3, respectively. The smallest observed NHANES III  $\hat{d}_{w0}$  values are somewhat more consistent with the simulated quantiles for Cases 4 through 6, although the  $\hat{d}_{w0} = 1.0$  for systolic blood pressure still falls below the simulated 0.005 quantile for Cases 4 and 5, and is approximately equal to the simulated 0.025 quantile for Case 6.

In addition, note that the three largest observed  $\hat{d}_{w0}$  values (for the hypertension indicator, the total cholesterol measure, and diastolic blood pressure) fall above the simulated upper 0.995 quantiles for each of cases 4 through 6. This, in conjunction with the abovementioned results for Cases 1 through 3, indicates that the twelve observed  $\hat{d}_{w0}$  are consistent with settings that produce substantially different true  $d_{w0}$  values for different variables.

Taken together, these simulation results suggest that for the twelve NHANES III variables examined, the stability of  $\hat{V}_{wh}$  may be substantially worse than one would anticipate from a simple count of SSUs within each stratum; and that the true stability measures  $d_{w0}$  may vary substantially from one variable to the next.

### 5.2.4 Diagnostic Plots

In a purely numerical sense,  $\hat{d}_{w0}$  depends on the magnitudes of the  $\tilde{V}(\hat{V}_{wh})$  relative to the terms  $2\hat{V}_{wh}^2$ . Consequently, diagnostic plots of  $\tilde{V}(\hat{V}_{wh})^{1/2}$  against  $\hat{V}_{wh}$  are useful in the identification of specific patterns and "problem strata" that lead to unusually high or low  $\hat{d}_{w0}$ .

Figures 1 through 3 give plots for the variables HAE2 (diagnosed hypertension), log(blood lead), and blood lead, respectively. Each plot was constructed with horizontal and vertical axes on the same scale. The plot for HAE2 has the bulk of its points well below a line with slope = 1 and intercept = 0. In addition, the values of  $\tilde{V}(\hat{V}_{wh})^{1/2}$  that are large in an absolute sense are still substantially less than the

corresponding  $\hat{V}_{wh}$ . This is consistent with the relatively large degrees-of-freedom value  $\hat{d}_{w0} = 23.7$ . The plot for log(blood lead) shows a somewhat greater concentration of points near the line with slope = 1 and intercept = 0, which is consistent with the somewhat smaller value  $\hat{d}_{w0} = 10.5$ .

The plot for blood lead shows one apparent outlier: the largest value of  $\tilde{V}(\hat{V}_{wh})^{1/2}$  is approximately equal to the corresponding  $\hat{V}_{wh}$ . For this stratum, we examined the terms  $\hat{V}_{wh}$  and  $p_{hi}^{-2} \hat{\sigma}_{2hi}^2$  for unusual patterns, e.g., extreme individual

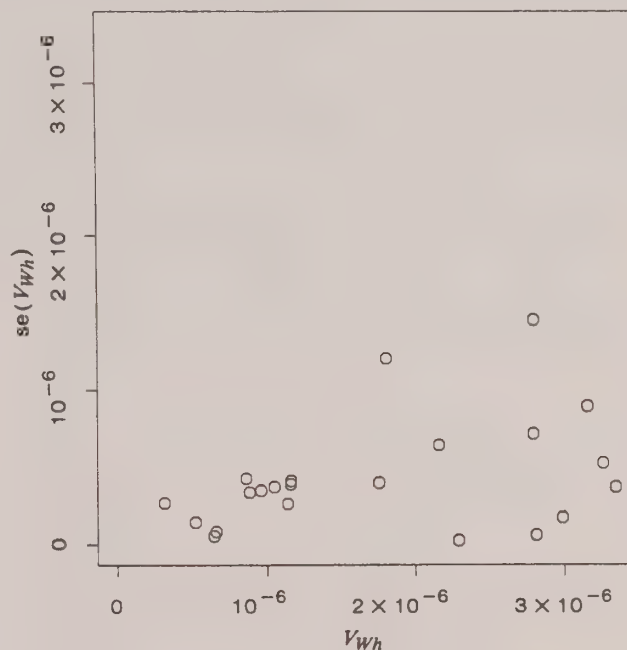


Figure 1. Plot of  $\tilde{V}(\hat{V}_{wh})^{1/2}$  against  $\hat{V}_{wh}$  for HAE2

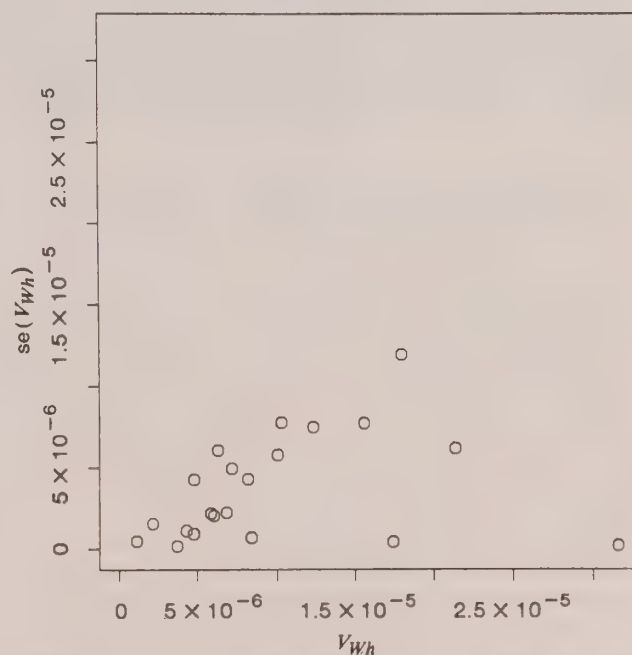


Figure 2. Plot of  $\tilde{V}(\hat{V}_{wh})^{1/2}$  against  $\hat{V}_{wh}$  for log (blood lead)

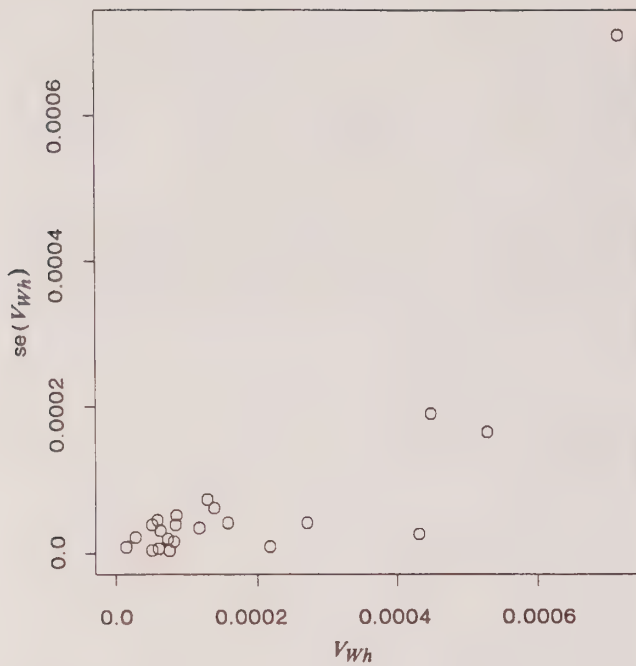


Figure 3. Plot of  $\tilde{V}(\hat{V}_{wh})^{1/2}$  against  $\hat{V}_{wh}$  for blood lead

values or extreme element-level weights. Here, one of the two associated  $p_{hi}^{-2} \hat{\sigma}_{2hi}^2$  values was approximately equal to zero and the other was the largest of all the PSU-level terms  $p_{hi}^{-2} \hat{\sigma}_{2hi}^2$ . In addition, the stratum in question had the largest  $\hat{V}_h$  value. However, this stratum did not display outlying values of  $\tilde{V}(\hat{V}_{wh})^{1/2}$  and  $\hat{V}_h$  for other related variables, *e.g.*, log (blood lead). Thus, the unusual pattern observed for blood lead may be attributable to a few very high observed values for the blood lead variable, rather than to the sample design or weighting as such. Within this context, note that at the population level in the U.S., lead measurements tend to have a roughly lognormal distribution, and high lead measurements show some tendency to be clustered together due to environmental factors.

### 5.3 Between-PSU Variance Estimates and the Variance Ratio $\hat{R}_{wv}$

Table 5 presents the estimates  $\hat{V}_B$  and  $\hat{R}_{wv}$ , and associated standard errors, for the twelve NHANES III variables. Of special interest are the columns labeled  $\tilde{V}(\hat{V}_B)^{-1} \tilde{V}(\hat{V}_w)$ , the proportion of the variance estimate  $\tilde{V}(\hat{V}_B)$  that is attributable to the within-PSU variance term; and  $\tilde{V}(\hat{R}_{wv})^{-1} \hat{V}_w^{-2} \hat{R}_{wv}^2 \tilde{V}(\hat{V}_w)$ , the corresponding proportion for  $\hat{R}_{wv}$ . Relatively large values for these proportions indicate that  $\tilde{V}(\hat{V}_w)$  makes a substantial contribution to  $\tilde{V}(\hat{V}_B)$  and  $\tilde{V}(\hat{R}_{wv})$  for the variables in question.

Note that the proportion  $\tilde{V}(\hat{R}_{wv})^{-1} \hat{V}_w^{-2} \hat{R}_{wv}^2 \tilde{V}(\hat{V}_w)$  is greater than or equal to 0.3 for blood lead, BP1K1 (systolic blood pressure) and BP1K5 (diastolic blood pressure). For blood lead and BP1K1, the large proportions arise primarily because of the relatively large value of  $\tilde{V}(\hat{V}_w)$ . For BP1K5,

Table 5

Estimates of  $\hat{V}_B$  and  $\hat{R}_{wv}$  for Twelve NHANES III Variables with Associated Standard Errors and Relative Within-PSU Contributions

Variable name	$\hat{V}_B$	se( $\hat{V}_B$ )	$\tilde{V}(\hat{V}_B)^{-1} \tilde{V}(\hat{V}_w)$
HAE2	0.0000126	0.0000188	0.020
HAE7	0.0000532	0.0000445	0.030
HAD1	-0.00000208	0.00000246	0.186
HAR3	0.0000825	0.0000703	0.047
BMPHT	0.0193	0.0114	0.027
BMPWT	0.0174	0.0400	0.096
HDRESULT	0.0887	0.0744	0.010
TCRESULT	0.270	0.253	0.031
LEAD	0.00269	0.00188	0.168
log(LEAD)	0.000468	0.000205	0.012
BP1K1	1.823	0.997	0.081
BP1K5	-0.0351	0.0793	0.367

	$\hat{R}_{wv}$	se( $\hat{R}_{wv}$ )	$\tilde{V}(\hat{R}_{wv})^{-1} \hat{V}_w^{-2} \hat{R}_{wv}^2 \tilde{V}(\hat{V}_w)$
HAE2	1.327	0.491	0.034
HAE7	1.648	0.556	0.077
HAD1	0.783	0.247	0.123
HAR3	1.676	0.600	0.122
BMPHT	1.864	0.530	0.089
BMPWT	1.168	0.391	0.126
HDRESULT	2.193	1.020	0.047
TCRESULT	1.458	0.436	0.063
LEAD	1.694	0.555	0.367
log(LEAD)	3.221	1.025	0.112
BP1K1	2.699	1.142	0.391
BP1K5	0.861	0.300	0.300

$\tilde{V}(\hat{V}_w)$  is not as large on a relative scale, but the proportion  $\tilde{V}(\hat{R}_{wv})^{-1} \hat{V}_w^{-2} \hat{R}_{wv}^2 \tilde{V}(\hat{V}_w)$  is still large because  $\hat{V}_w$  is not small relative to  $\tilde{V}(\hat{Y})$ . For all three variables, the relatively large values of  $\tilde{V}(\hat{R}_{wv})^{-1} \hat{V}_w^{-2} \hat{R}_{wv}^2 \tilde{V}(\hat{V}_w)$  indicate that it is important to account for the variance  $V(\hat{V}_w)$  when one considers the stability of  $\hat{R}_{wv}$ . For BP1K5, a similar comment applies to the effect of  $V(\hat{V}_w)$  on the stability of  $\hat{V}_B$ .

## 6. DISCUSSION

This paper has presented three main ideas. First, due to the role that estimated within-PSU variances  $\hat{V}_{wh}$  play in survey design and analysis, it is important to account for the sampling error encountered in estimation of  $V_{wh}$ . Second, standard design-based estimation methods lead to relatively simple estimators of the design variance of  $\hat{V}_{wh}$ . In general, interpretation of these stability measures



requires some caution. However, they can provide useful diagnostics for the identification of variables for which the instability of  $\hat{V}_{wh}$  is especially problematic, or has an especially pronounced effect on the variance of related quantities like  $\hat{V}_B$  and  $\hat{R}_{wv}$ . Third, the application to the U.S. Third National Health and Nutrition Examination Survey (NHANES III), and associated simulation work, indicated the following.

- (i) For different sets of variables, the observed stability measures  $\hat{d}_{w0}$  are consistent with substantially different sets of stability conditions.
- (ii) For some variables, the estimators  $\hat{V}_{wh}$  are considerably less stable than one would anticipate from a direct count of secondary sample units.
- (iii) For some variables, the estimated variance of  $\hat{V}_{wh}$  makes a substantial contribution to the estimated variances of the estimated between-PSU variance  $\hat{V}_B$  and the variance ratio  $\hat{R}_{wv}$ .

### ACKNOWLEDGEMENTS

We thank Van Parsons, Cliff Johnson and other NCHS statisticians for providing access to the NHANES III dataset; and for sharing a wealth of information regarding the NHANES III – Phase I project. We also thank Van Parsons and two anonymous referees for helpful comments on earlier versions of this paper. This research was supported in part by the National Center for Health Statistics. The views expressed in this paper are those of the authors and do not necessarily represent the policies of the National Center for Health Statistics.

### REFERENCES

- FULLER, W.A. (1984). Least squares and related analyses for complex survey design. *Survey Methodology*, 10, 97-118.
- HANSEN, M.H., HURWITZ, W.N., and MADOW, W.G. (1953). *Sample Survey Methods and Theory, Volume I: Methods and Applications*. New York: John Wiley.
- HERZOG, T.N., and SCHEUREN, F.J. (1976). Dallying with some CPS design effects for proportions. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 396-401.
- JANG, D.S., and ELTINGE, J.L. (1996). Use of Within-PSU Variances and Errors-in-Variables Regression to Assess the Stability of a Standard Design-Based Variance Estimator. Unpublished manuscript, Department of Statistics, Texas A&M University.
- KORN, E.I., and GRAUBARD, B.G. (1995). Analysis of large health surveys: accounting for the sampling design. *Journal of the Royal Statistical Society, Series A*, 158, 263-295.
- NATIONAL CENTER FOR HEALTH STATISTICS (1996). National Health and Nutrition Examination Survey III Report (in press). National Center for Health Statistics, Hyattsville, MD.
- RUST, K., and KALTON, G. (1987). Strategies for collapsing strata for variance estimation. *Journal of Official Statistics*, 3, 69-81.
- SATTERTHWAITE, F.E. (1941). Synthesis of variance. *Psychometrika*, 6, 309-316.
- SATTERTHWAITE, F.E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2, 110-114.
- WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.





# Asymptotic Variance for Sequential Sampling Without Replacement With Unequal Probabilities

YVES G. BERGER<sup>1</sup>

## ABSTRACT

We propose a second-order inclusion probability approximation for the Chao plan (1982) to obtain an approximate variance estimator for the Horvitz and Thompson estimator. We will then compare this variance with other approximations provided for the randomized systematic sampling plan (Hartley and Rao 1962), the rejective sampling plan (Hájek 1964) and the Rao-Sampford sampling plan (Rao 1965 and Sampford 1967). Our conclusion will be that these approximations are equivalent if the first-order inclusion probabilities are small and if the size of the sample is large.

**KEY WORDS:** Sampling with replacement; Randomized systematic sampling plan; Rejective sampling plan; Rao-Sampford sampling plan; Inclusion probabilities; Horvitz-Thompson; Yates-Grundy.

## 1. INTRODUCTION

Consider a finite population  $U_N$  containing  $N$  units and a subset  $U_k$  of  $U_N$  comprising the first units  $k$  of  $U_N$ . Let  $\pi_{(k;i)}$  denote the first-order inclusion probabilities for a population  $U_k$ . We assume that they are proportional to an auxiliary variable. These probabilities have two arguments: the size  $k$  of the population and the serial number  $i$  of the unit within the population. We assume that  $\pi_{(k;i)} < 1$  for all  $i$  and that all  $k > n$ . This hypothesis has more chance of breaking down when  $k$  is small, i.e., close to  $n$ . We can solve this problem by assuming that the values of the auxiliary variable show little dispersion for those units occurring at the beginning of the population.

Let  $\pi_{(k;i,j)}$  denote the second-order inclusion probability of units  $i$  and  $j$  for a population  $U_k$ . These probabilities are dependent on the sampling plan used.

We will use the Horvitz-Thompson estimator (1951) to estimate the total  $\sum_{i=1}^N Y_i$  of a variable  $Y$ . This estimator is given by

$$t_{HT} = \sum_{i \in S_N} \frac{Y_i}{\pi_{(N;i)}}; \quad (1)$$

where  $S_N$  is a sample of  $U_N$ . We assume that the size of  $S_N$  is constant and equal to  $n$ .

Given that the size of the sample is fixed, a variance estimator of (1) is given by the Yates-Grundy estimator (1953),

$$\hat{V} = \sum_{j \in S_N} \sum_{i \in S_N; i < j} \frac{-\Delta_{(N;i,j)}}{\pi_{(N;i,j)}} \left[ \frac{Y_i}{\pi_{(N;i)}} - \frac{Y_j}{\pi_{(N;j)}} \right]^2, \quad (2)$$

where

$$\Delta_{(N;i,j)} = \pi_{(N;i,j)} - \pi_{(N;i)}\pi_{(N;j)}. \quad (3)$$

Let us consider the sample size sequence  $\{n_1, n_2, \dots, n_v, \dots\}$  and the population size sequence  $\{N_1, N_2, \dots, N_v, \dots\}$ , where  $n_v$  and  $N_v$  increase whenever  $v \rightarrow \infty$ . To simplify the problem we eliminate the index  $v$ .

The asymptotic approach used here is that of Hájek (1964):

$$d = \sum_{j=1}^N \pi_{(N;j)} [1 - \pi_{(N;j)}] \rightarrow \infty,$$

which means that  $n \rightarrow \infty$  and  $(N - n) \rightarrow \infty$ , given that  $d \leq \sum_{j=1}^N [1 - \pi_{(N;j)}] = N - n$  and that  $d \leq \sum_{j=1}^N \pi_{(N;j)} = n$ .

In section 2, we introduce the Chao sampling plan (1982) as well as three results linked to first and second-order inclusion probabilities. In section 3, we provide an approximation of  $\pi_{(N;i,j)}$ . In section 4, we propose an approximation of the Yates-Grundy variance. Section 5 compares this variance approximation with other approximations proposed for the randomized systematic plan, the rejective plan and the Rao-Sampford plan. Two numerical examples are provided in section 6.

## 2. CHAO SAMPLING PLAN

This is a sampling plan without replacement with unequal probabilities, of fixed size. This method is a generalization of the method used by McLeod and Bellhouse (1983) for a simple plan.

Let  $S_k$  denote a sample of size  $n$  of  $U_k$  with a set  $\{\pi_{(k;i)}; i \in U_k\}$  of first-order inclusion probabilities. The Chao plan provides for a sample  $S_{k+1}$  of size  $n$  of  $U_{k+1}$  with a set  $\{\pi_{(k+1;i)}; i \in U_{k+1}\}$  of first-order inclusion probabilities. The method entails selecting the  $(k+1)$ -th unit with the probability  $\pi_{(k+1;k+1)}$ . If this unit is not selected, then we take  $S_{k+1} = S_k$ ; otherwise we take  $S_{k+1} = S_k \cup \{k+1\} \setminus \{j\}$ , where  $j$  is a unit selected at random within  $S_k$ . The procedure starts from an initial sample  $S_n = U_n$  comprising the first units  $n$  of the population.

<sup>1</sup> Yves Berger, Université Libre de Bruxelles, Laboratoire de Méthodologie du Traitement des Données, C.P. 124, Avenue Jeanne, 44, Bruxelles, Belgique, E-Mail : yvberger@ulb.ac.be

The Chao plan provides the advantage of being sequential. In fact, it allows us to select a sample through a simple sequential run of the population. The systematic plan is another sequential plan that is often used. However, the latter is inconvenient in that it induces zero second-order inclusion probabilities. We can avoid this problem by randomizing the systematic plan. In such a case, the population is ordered at random before the sample is selected. This operation eliminates in part the problem of zero second-order inclusion probabilities. As will be seen at the end of this section, the Chao plan offers the advantage of not having any zero second-order inclusion probabilities. Randomization is therefore not needed for the latter.

The rejective plan and the Rao-Sampford plan are inconvenient in that they are not sequential. In fact, the units are selected at random with replacement within the population. If a unit is selected twice, we are forced to select a new sample. These two plans, although they are more easily understood, are more difficult to implement than the Chao plan.

The following theorem, which is a direct application of the theorem given by Chao (1982), provides a relation between the first-order inclusion probability  $\pi_{(k;i)}$  of the  $i$ -th unit of  $U_k$  and the first-order inclusion probability  $\pi_{(k+1;i)}$  of the  $i$ -th unit of  $U_{k+1}$ .

### Theorem 1

$$\pi_{(k+1;i)} = \begin{cases} [1 - \pi_{(k+1;k+1)} R_{(k;i)}] \pi_{(k;i)}, & \text{for } i < k + 1; \\ \pi_{(k+1;k+1)}, & \text{for } i = k + 1; \end{cases} \quad (4)$$

where

$$R_{(k;i)} = \begin{cases} \frac{1 - \pi_{(n+1;i)}}{\pi_{(n+1;n+1)}}, & \text{for } k = n, \\ \frac{1}{n}, & \text{for } k \geq n + 1. \end{cases} \quad (5)$$

The second-order inclusion probabilities can be calculated iteratively using the following theorem:

### Theorem 2 (Chao, 1982)

$$\pi_{(k;i,j)} = \begin{cases} \{1 - \pi_{(k;k)} [R_{(k-1;i)} + R_{(k-1;j)}]\} \pi_{(k-1;i,j)}, & \text{for } i < j < k; \\ \pi_{(k;k)} [1 - R_{(k-1;i)}] \pi_{(k-1;i)}, & \text{for } i < j = k. \end{cases}$$

Bethlehem and Schuerhoff (1984) give a sufficient and necessary condition for the second-order inclusion probabilities to be strictly positive for a population  $U_k$ :

$$\# \{i : i \leq \ell \text{ and } \pi_{(i;i)} = 1\} \neq n - 1, \text{ for } \ell \text{ such that } n < \ell \leq k.$$

Since  $\pi_{(i;i)} < 1$  for all  $i$  and  $\ell$  such that  $i \leq \ell \leq k$ , this condition is always met. Therefore, within the framework of this article, we will never have zero second-order inclusion probabilities.

Moreover, the quantity  $\Delta_{(N;i,j)}$  is always negative if we use the Chao plan (Chao 1982, p. 656). Then the Yates-Grundy variance offers the advantage of always being positive.

## 3. APPROXIMATION OF SECOND-ORDER INCLUSION PROBABILITIES

The following theorem provides us with an asymptotic expression for second-order inclusion probabilities for the Chao plan.

### Theorem 3

$$\pi_{(N;i,j)} \approx \begin{cases} \pi_{(N;i)} \pi_{(N;j)} \frac{n-1}{n-p_{(j)}}, & \text{if } j > n+1; \\ \pi_{(N;i)} \pi_{(N;j)} \frac{\pi_{(n+1;i)} + \pi_{(n+1;j)} - 1}{\pi_{(n+1;i)} \pi_{(n+1;j)}}, & \text{if } j \leq n+1; \end{cases} \quad (6)$$

where  $p_{(j)} = \pi_{(j;j)}$  and  $i < j$ .

The proof of this theorem can be found in Appendix I.

Note that this approximation has a different structure depending on whether  $j > n+1$  or  $j \leq n+1$ . To avoid this problem, we will use a plausible condition for the auxiliary variable so that these two structures will be equivalent. Let us consider the hypothesis given in the introduction, that the values of the auxiliary variable show little dispersion for the first units  $n+1$  of the population. More precisely, we assume that the auxiliary variable is constant for the first units  $n+1$ , i.e.:

$$\pi_{(n+1;i)} = \frac{n}{n+1} \quad \text{for } i \leq n+1.$$

In this case,

$$\frac{\pi_{(n+1;i)} + \pi_{(n+1;j)} - 1}{\pi_{(n+1;i)} \pi_{(n+1;j)}} = \frac{n-1}{n - \pi_{(n+1;j)}}.$$

By using (6), we have the following approximation for second-order inclusion probabilities

$$\pi_{(N;i,j)} \approx \pi_{(N;i)} \pi_{(N;j)} \frac{n-1}{n-p_{(j)}} \quad \text{if } i < j; \quad (7)$$

where

$$p_{(j)} = \begin{cases} \pi_{(j;j)}, & \text{if } j > n+1, \\ \pi_{(n+1;j)}, & \text{if } j \leq n+1. \end{cases} \quad (8)$$

## 4. VARIANCE ESTIMATOR

Relation (7) leads to the following approximation for  $\Delta_{(N;i,j)}$ :



$$\tilde{\Delta}_{(N;i,j)} = \pi_{(N;i)} \pi_{(N;j)} \frac{p_{(j)} - 1}{n - p_{(j)}}, \quad \text{if } i < j. \quad (9)$$

(2), (7) and (9) provide an asymptotic expression for the Yates-Grundy estimator.

$$\tilde{V}_C = \frac{1}{[n-1]} \sum_{j \in S_N} [1 - p_{(j)}] \sum_{i \in S_N, i < j} \left[ \frac{Y_i}{\pi_{(N;i)}} - \frac{Y_j}{\pi_{(N;j)}} \right]^2. \quad (10)$$

But this expression tends to underestimate the variance. In fact, to establish relation (6), we use approximation (19) from Appendix I. This approximation always implies that:

$$\pi_{(N;i,j)} < \pi_{(N;i)} \pi_{(N;j)} \frac{n-1}{n - p_{(j)}}. \quad (11)$$

This can easily be verified if we observe that (20) is obtained from (18) using approximation (19). Inequality (11) is therefore true for  $j > n+1$ . For  $j \leq n+1$ , it is sufficient to observe that (21) is also obtained from (19). Inequality (11) implies that:

$$\frac{-\Delta_{(N;i,j)}}{\pi_{(N;i,j)}} > \frac{1 - p_{(j)}}{n-1}, \quad (12)$$

given that  $\Delta_{(N;i,j)} < 0$ . From (2), (10) and (12), we have effectively

$$\hat{V} > \tilde{V}_C.$$

To overcome this problem of variance underestimation, we plan to make an adjustment on (9). It is well known that:

$$\sum_{i=1}^N \pi_{(N;i,j)} = (n-1) \pi_{(N;j)}. \quad (13)$$

Approximation (7) does not abide by constraint (13). The adjustment involves assuming that the  $p_{(j)}$  are unknown and selecting them so as to satisfy (13) for the second-order probability approximation, *i.e.*:

$$\sum_{i=1}^{j-1} \pi_{(N;i)} \pi_{(N;j)} \frac{n-1}{n - p_{(j)}} + \sum_{i=j+1}^N \pi_{(N;i)} \pi_{(N;j)} \frac{n-1}{n - p_{(i)}} = (n-1) \pi_{(N;j)}.$$

This constraint can be written as follows:

$$\sum_{i=1}^{j-1} \pi_{(N;i)} + \sum_{i=j+1}^N \pi_{(N;i)} \frac{n - p_{(j)}}{n - p_{(i)}} = n - p_{(j)}. \quad (14)$$

Given that  $\sum_{j=1}^N \pi_{(N;j)} = n$ , constraint (14) is practically verified if

$$p_{(i)} = \pi_{(N;i)} \quad (15)$$

$$\sum_{i=j+1}^N \pi_{(N;i)} \frac{n - \pi_{(N;j)}}{n - \pi_{(N;i)}} \approx \sum_{i=j+1}^N \pi_{(N;i)}. \quad (16)$$

Relation (16) is plausible given that the difference between the left and right sides of (16) has as its lower bound

$$\frac{1}{n} \sum_{i=j+1}^N \pi_{(N;i)} [\pi_{(N;i)} - \pi_{(N;j)}],$$

and as its upper bound

$$\frac{1}{n-1} \sum_{i=j+1}^N \pi_{(N;i)} [\pi_{(N;i)} - \pi_{(N;j)}].$$

These two bounds are close to zero when the  $\pi_{(N;i)}$  show little dispersion. This means that solution (15) is appropriate when the  $\pi_{(N;j)}$  are small. Furthermore, the greater the value of  $j$ , the closer the two bounds are to zero. Therefore, solution (15) verifies (13) all the more as  $j$  is large. This implies that our approximation (9) is very good for the duplicate pairs  $(i, j)$  ( $i < j$ ) such that the unit  $j$  is located at the end of the population. In fact, we want approximation (9) to be the best for the duplicate pairs  $(i, j)$  whose presence in the sample is highly probable (*i.e.*, for the pairs  $(i, j)$  ( $i < j$ ) for which  $\pi_{(N;j)}$  is the largest). It is therefore preferable to place the units having high first-order inclusion probabilities at the end of the population.

If we choose to have  $p_{(i)} = \pi_{(N;i)}$ , we have  $p_{(i)}$  smaller than (8). This leads to a larger variance approximation. This solution is all the more acceptable as it corresponds to the result of the simple plan without replacement. In fact, if we replace within (7)  $\pi_{(N;i)}$ ,  $\pi_{(N;j)}$  and  $p_{(j)}$  by  $n/N$ , we obtain

$$\pi_{(N;i,j)} \approx \frac{n(n-1)}{N(N-1)}, \quad \text{if } i > n+1.$$

This expression corresponds, quite clearly, to the result of the simple plan without replacement.

In conclusion, we approximate  $\Delta_{(N;i,j)}$  through (9) with  $p_{(i)} = \pi_{(N;i)}$ . We assume that the population is ordered in such a way that the units having small  $\pi_{(N;i)}$  are located at the beginning of the population and that the units having large  $\pi_{(N;j)}$  are located at the end of the population. We also assume that the  $\pi_{(N;i)}$  do not show too much dispersion for the first units  $n+1$  of the population.

## 5. COMPARISON WITH OTHER PLANS

Instead of comparing the second-order inclusion probabilities, we will compare the quantities  $-\Delta_{(N;i,j)}/\pi_{(N;i,j)}$  which are of some use in calculating the Yates-Grundy variance. We will examine what these quantities provide for the Chao plan, the randomized systematic plan (Hartley and Rao 1962), the rejective plan (Hájek 1964) and the Rao-Sampford plan (Rao 1965, and Sampford 1967).

**Theorem 4**

$$\frac{-\Delta_{(N;i,j)}}{\pi_{(N;i,j)}} \approx \begin{cases} \frac{1 - \pi_{(N;j)}}{n - 1}, & \text{For the Chao plan;} \\ \frac{1 - \pi_{(N;i)} - \pi_{(N;j)}}{n - 1}, & \text{for the randomized systematic plan;} \\ \frac{n[1 - \pi_{(N;i)}][1 - \pi_{(N;j)}]}{d(n - 1)}, & \text{for the rejective plan and the Rao-Sampford plan.} \end{cases}$$

The proof of this theorem can be found in Appendix II.

It is important to note that the proposed approximation for the randomized systematic plan comes from Deville's approximation (p. 21) and not from the famous Hartley-Rao approximation (1962). We were not able to use the Hartley-Rao formula because the latter is based on the asymptotic hypothesis,  $n$  fixed and  $N \rightarrow \infty$ , which is different from that adopted in this paper.

We observe that if the  $\pi_{(N;i)}$  are small,  $-\Delta_{(N;i,j)}/\pi_{(N;i,j)}$  is equivalent for the Chao plan and for the systematic plan. However, we observe that  $-\Delta_{(N;i,j)}/\pi_{(N;i,j)}$  is always smaller in the systematic case than it is in the Chao case. This is certainly due to the fact that the approximation for the systematic plan underestimates  $-\Delta_{(N;i,j)}/\pi_{(N;i,j)}$ . This can be confirmed by replacing  $\pi_{(N;i)}$  and  $\pi_{(N;j)}$  by  $n/N$ . We then have

$$\frac{-\Delta_{(N;i,j)}}{\pi_{(N;i,j)}} \approx \frac{N - 2n}{N(n - 1)},$$

for the randomized systematic plan. This is equivalent to a simple plan, thus

$$\frac{-\Delta_{(N;i,j)}}{\pi_{(N;i,j)}} = \frac{N - n}{N(n - 1)}.$$

We intend to adjust the approximation of  $-\Delta_{(N;i,j)}/\pi_{(N;i,j)}$  for the systematic plan by multiplying it by

$$\frac{N - n}{N - 2n} = \frac{1 - f}{1 - 2f},$$

where  $f = n/N$  is the sampling rate.

The approximation of  $-\Delta_{(N;i,j)}/\pi_{(N;i,j)}$  for the Chao plan is also of the same magnitude as that of the rejective plan. In fact, if the  $\pi_{(N;i)}$  are small, we have the approximation

$$\frac{n[1 - \pi_{(N;i)}]}{d} \approx \frac{n[1 - \pi_{(N;i)}]}{[1 - \pi_{(N;i)}] \sum_{j=1}^N \pi_{(N;j)}} \approx 1.$$

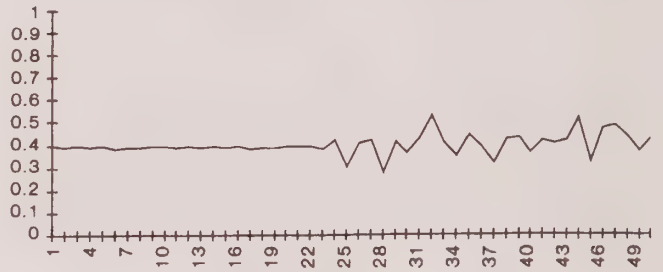
Therefore, the Yates-Grundy estimator is approximately the same whether we use the Chao plan, the randomized systematic plan, the rejective plan or the Rao-Sampford plan, for large  $n$  and small  $\pi_{(N;i)}$ .

**6. NUMERICAL EXAMPLES**

The two following examples correspond to two extreme cases. In the first example, the  $\pi_{(N;i)}$  show little dispersion; in the second, they show much more dispersion. Let us consider a small sample of size 20. The population size is 50 so that the  $\pi_{(N;i)}$  are not too small. We have willingly opted for a bad situation in order to show that even with a sample of size 20 and a small population, the asymptotic results nevertheless represent a good approximation.

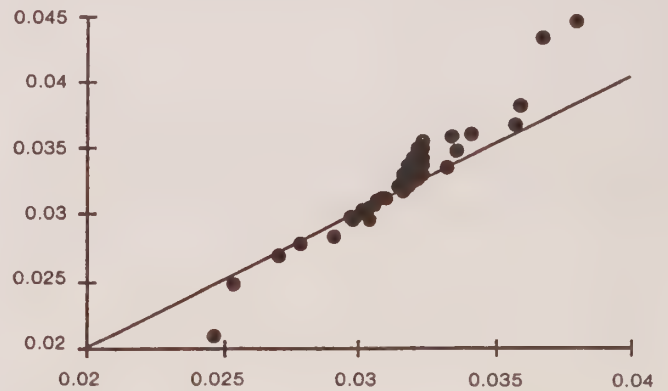
**Example 1**

Let us consider the first-order inclusion probabilities represented in Figure 1.



**Figure 1.** First-order inclusion probabilities in the case of Example 1

Figure 2 shows, on the Y axis, the true values of  $-\Delta_{(N;i,j)}/\pi_{(N;i,j)}$  for the Chao plan and, on the X axis, the approximations. We have also represented the straight line where the approximations are equal to the true values. The approximations are all the better as the points are close to the straight line.



**Figure 2.** Approximations and true values of  $-\Delta_{(N;i,j)}/\pi_{(N;i,j)}$ , in the case of Example 1

We have a mean error of  $-0.000569$  with a standard deviation of  $0.0015996$ . This is very small in relation to the order of magnitude of the approximations. The centre of gravity of the scatter plot is located in  $(0.0313; 0.0318)$ . It might seem surprising that there are less points at the left of the centre of gravity than at the right. This is simply due to the fact that most of the points at the left of the centre of gravity overlap.



We observe that the pairs  $(i, j)$  with  $i < j$  such that  $\pi_{(N;i,j)}$  is large correspond to points located on the left. They are the pairs showing the best approximation. Moreover, there is a high probability that these pairs are located within the sample given that  $\pi_{(N;i,j)}$  is large. Therefore, our approximate variance (10) is definitely acceptable.

### Example 2

The first-order inclusion probabilities are given in Figure 3. Here we notice that these probabilities are more dispersed than in Example 1. Figure 4 provides the true values as well as the approximations of  $-\Delta_{(N;i,j)}/\pi_{(N;i,j)}$ .

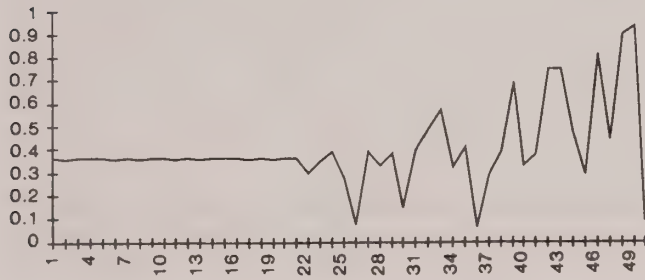


Figure 3. First-order inclusion probabilities in the case of Example 2

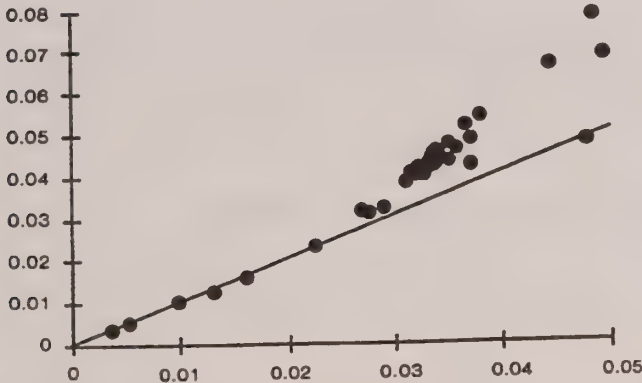


Figure 4. Approximations and true values of  $-\Delta_{(N;i,j)}/\pi_{(N;i,j)}$ , in the case of Example 2

We have a mean error of  $-0.006999$  with a standard deviation of  $0.006438$ . The centre of gravity of the scatter plot is located in  $(0.02957; 0.036606)$ .

We reach the same conclusion as in Example 1. The second example leads to worse approximations. This is simply due to the high first-order inclusion probabilities.

## 7. CONCLUSION

The Chao plan provides a number of advantages: (i) it is sequential; (ii) the second-order inclusion probabilities are positive; and (iii) the Yates-Grundy variance is always positive. On the other hand, the second-order inclusion probabilities are difficult to calculate. That is why we propose

to approximate them. We have observed that this approximation is better when the beginning of the population consists of units having small  $\pi_{(N;i)}$  and the end of the population consists of units having large  $\pi_{(N;i)}$ . We have compared our approximation with other approximations provided for the randomized systematic plan, the rejective plan and the Rao-Sampford plan. We have concluded that these approximations are equivalent if the first-order inclusion probabilities are small and if the size of the sample is large. The two numerical examples which close this paper confirm the sound results of our approximation.

## APPENDIX I

### Proof of Theorem 3

Before proving this theorem, we will demonstrate the following two lemmas.

#### Lemma 1

$$\pi_{(k;i)} = p_{(i)}^* \prod_{\ell=a_i^*}^k \left[ 1 - \pi_{(\ell;\ell)} \frac{1}{n} \right];$$

where

$$p_{(i)}^* = \begin{cases} \pi_{(i;i)} & \text{if } i > n+1; \\ \pi_{(n+1;i)} & \text{if } i \leq n+1; \end{cases}$$

$$a_i^* = \begin{cases} i+1 & \text{if } i > n+1; \\ n+2 & \text{if } i \leq n+1. \end{cases} \quad (17)$$

#### Lemma 2

$$\pi_{(k;i,j)} = q_{(j)}^* \prod_{\ell=a_j^*}^k \left[ 1 - \pi_{(\ell;\ell)} \frac{2}{n} \right];$$

where  $i < j$ ,

$$q_{(j)}^* = \begin{cases} \pi_{(j-1;i)} \pi_{(j;j)} \left( 1 - \frac{1}{n} \right) & \text{if } j > n+1; \\ \pi_{(n+1;i)} + \pi_{(n+1;j)} - 1 & \text{if } j \leq n+1; \end{cases}$$

and  $a_j^*$  is defined by (17).

Now, with these two lemmas, we can demonstrate Theorem 3.

### Proof of Theorem 3

**Case 1:** If  $j > n+1$ , using Lemma 2, we have

$$\pi_{(N;i,j)} = \pi_{(j-1;i)} \pi_{(j;j)} \left( 1 - \frac{1}{n} \right) \prod_{\ell=j+1}^N \left[ 1 - \pi_{(\ell;\ell)} \frac{2}{n} \right].$$

On the basis of Lemma 1, this last expression becomes

$$\pi(N; i, j) = p_{(i)}^* \pi_{(j; j)} \left(1 - \frac{1}{n}\right) \prod_{\ell=a_i}^{j-1} \left[1 - \pi_{(\ell; \ell)} \frac{1}{n}\right] \prod_{q=j+1}^N \left[1 - \pi_{(q; q)} \frac{2}{n}\right].$$

By multiplying this last expression by

$$\left[ \frac{1 - \pi_{(j; j)} \frac{1}{n}}{1 - \pi_{(j; j)} \frac{1}{n}} \right] \prod_{\ell=j+1}^N \left[ \frac{1 - \pi_{(\ell; \ell)} \frac{1}{n}}{1 - \pi_{(\ell; \ell)} \frac{1}{n}} \right] \left[ \frac{1 - \pi_{(\ell; \ell)} \frac{2}{n}}{1 - \pi_{(\ell; \ell)} \frac{2}{n}} \right] = 1,$$

and by regrouping certain terms, we obtain

$$\pi_{(N; i, j)} = \pi_{(j; j)} p_{(i)}^* \left[ \frac{n-1}{n - \pi_{(j; j)}} \right] \prod_{\ell=a_i}^N \left[ \frac{1 - \pi_{(q; q)} \frac{2}{n}}{1 - \pi_{(q; q)} \frac{1}{n}} \right].$$

On the basis of Lemma 1, this last expression becomes

$$\pi_{(N; i, j)} = \left[ \frac{n-1}{n - \pi_{(j; j)}} \right] \pi_{(N; i)} \pi_{(j; j)} \prod_{\ell=j+1}^N \left[ \frac{1 - \pi_{(\ell; \ell)} \frac{2}{n}}{1 - \pi_{(\ell; \ell)} \frac{1}{n}} \right]. \quad (18)$$

If  $n$  is sufficiently large

$$\begin{aligned} \frac{1 - \pi_{(\ell; \ell)} \frac{2}{n}}{1 - \pi_{(\ell; \ell)} \frac{1}{n}} &\approx \left[1 - \pi_{(\ell; \ell)} \frac{2}{n}\right] \left[1 + \pi_{(\ell; \ell)} \frac{1}{n}\right]; \\ &\approx 1 + \frac{\pi_{(\ell; \ell)}}{n} - \frac{2\pi_{(\ell; \ell)}}{n} - \frac{2\pi_{(\ell; \ell)}^2}{n^2}; \\ &\approx 1 - \frac{\pi_{(\ell; \ell)}}{n}. \end{aligned} \quad (19)$$

Then (18) becomes,

$$\pi_{(N; i, j)} \approx \left[ \frac{n-1}{n - \pi_{(j; j)}} \right] \pi_{(N; i)} \pi_{(j; j)} \prod_{\ell=j+1}^N \left[1 - \pi_{(\ell; \ell)} \frac{1}{n}\right]. \quad (20)$$

Finally, on the basis of Lemma 1, this last expression can be written:

$$\pi_{(N; i, j)} \approx \pi_{(N; i)} \pi_{(N; j)} \frac{n-1}{n - \pi_{(j; j)}}.$$

**Case 2:** If  $j \leq n+1$ , Lemma 2 provides

$$\pi_{(N; i, j)} = [\pi_{(n+1; i)} + \pi_{(n+1; j)} - 1] \prod_{\ell=n+2}^N \left[1 - \pi_{(\ell; \ell)} \frac{2}{n}\right];$$

in other words

$$\pi_{(N; i, j)} = \prod_{\ell=n+2}^N \left[1 - \pi_{(\ell; \ell)} \frac{1}{n}\right] \prod_{q=n+2}^N \left[ \frac{1 - \pi_{(q; q)} \frac{2}{n}}{1 - \pi_{(q; q)} \frac{1}{n}} \right] [\pi_{(n+1; i)} + \pi_{(n+1; j)} - 1].$$

By using approximation (19), we obtain

$$\begin{aligned} \pi_{(N; i, j)} &\approx \left\{ \prod_{\ell=n+2}^N \left[1 - \pi_{(\ell; \ell)} \frac{1}{n}\right] \right\}^2 \\ &\quad \frac{\pi_{(n+1; i)} \pi_{(n+1; j)}}{\pi_{(n+1; i)} \pi_{(n+1; j)}}. \end{aligned}$$

On the basis of Lemma 1, we obtain finally

$$\pi_{(N; i, j)} \approx \pi_{(N; i)} \pi_{(N; j)} \frac{\pi_{(n+1; i)} + \pi_{(n+1; j)} - 1}{\pi_{(n+1; i)} \pi_{(n+1; j)}}. \quad (21)$$

Q.E.D.

## APPENDIX II

### Proof of Theorem 4

- For the Chao plan, it is sufficient to use (6), (9) and (15).
- For the randomized systematic plan, it is sufficient to use the approximation of the  $\pi_{(N; i, j)}$  given by Deville (p. 21)

$$\pi_{(N; i, j)} \approx \pi_{(N; i)} \pi_{(N; j)} \frac{n-1}{n - \pi_{(N; i)} - \pi_{(N; j)}}. \quad (22)$$

This expression is obtained from the hypothesis

$$\text{Max}_{1 \leq i \leq N} \left\{ \frac{\pi_{(N; i)}}{n} \right\} \rightarrow 0.$$

This last hypothesis is verified since  $n \rightarrow \infty$ .



- For the rejective plan, using Hájek's result (1964, p. 1508), we have

$$\frac{-\Delta_{(N;i,j)}}{\pi_{(N;i,j)}} \approx \frac{[1 - \pi_{(N;i)}][1 - \pi_{(N;j)}]}{d - [1 - \pi_{(N;i)}][1 - \pi_{(N;j)}]}, \quad (23)$$

for  $d \rightarrow \infty$ . We note that (23) remains valid for the Rao-Sampford plan (see Hájek 1981, Theorem 8.2, p. 82). Using the approximation (Hájek 1964, p. 1521),

$$\{d - [1 - \pi_{(N;i)}][1 - \pi_{(N;j)}]\}^{-1} \approx \frac{n}{d(n-1)},$$

we obtain the result of the theorem.

Q.E.D.

### ACKNOWLEDGEMENTS

The author wishes to thank the referees who submitted a number of constructive comments that led to considerable improvements.

### REFERENCES

- BETHLEHEM, J.G., and SCHUERHOFF, H. (1984). Second-order inclusion probabilities in sequential sampling without replacement with unequal probabilities. *Biometrika*, 71, 642-644.
- CHAO, M.T. (1982). A general purpose unequal probability sampling plan. *Biometrika*, 69, 653-656.
- DEVILLE, J.-C. (No date). Cours de sondage, Chapitre III: les outils de bases. Lecture notes, ENSAE, Paris.
- HÁJEK, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*, 35, 1491-1523.
- HÁJEK, J. (1981). *Sampling from a Finite Population*. New York and Basel: Marcel Dekker, Inc.
- HARTLEY, H.O., and RAO, J.N.K. (1962). Sampling with unequal probabilities without replacement. *Annals of Mathematical Statistics*, 33, 350-374.
- HORVITZ, D.G., and THOMPSON, D.J. (1951). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.
- MCLEOD, A.I., and BELLHOUSE, D.R. (1983). A convenient algorithm for drawing a simple random sample. *Applied Statistics*, 32, 2.
- RAO, J.N.K. (1965). On two simple schemes of unequal probability sampling without replacement. *Journal of the Indian Statistical Association*, 3, 173-180.
- SAMPFORD, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54, 494-513.
- YATES, F., and GRUNDY, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society, series B*, 1, 253-261.





# Applications of Spatial Smoothing to Survey Data

ANN COWLING, RAY CHAMBERS, RAY LINDSAY and BHAMATHY PARAMESWARAN<sup>1</sup>

## ABSTRACT

In this paper we present two applications of spatial smoothing using data collected in a large scale economic survey of Australian farms: one a small area and the other a large area application. In the small area application, we describe how the sample weights can be spatially smoothed in order to improve small area estimates. In the large area application, we give a method for spatially smoothing and then mapping the survey data. The standard method of weighting in the survey is a variant of linear regression weighting. For the small area application, this method is modified by introducing a constraint on the spatial variability of the weights. Results from a small scale empirical study indicate that this decreases the variance of the small area estimators as expected, but at the cost of an increase in their bias. In the large area application, we describe the nonparametric regression method used to spatially smooth the survey data as well as techniques for mapping this smoothed data using a Geographic Information System (GIS) package. We also present the results of a simulation study conducted to determine the most appropriate method and level of smoothing for use in the maps.

**KEY WORDS:** Kernel estimation; Mapping survey data; Small area estimation; Survey weighting.

## 1. INTRODUCTION

The Australian Bureau of Agricultural and Resource Economics (ABARE) is the applied economic research organisation attached to the Department of Primary Industries and Energy. Amongst its information gathering activities, ABARE conducts annual surveys of selected Australian agricultural industries which provide a broad range of information on the economic and physical characteristics of farm business units.

The largest survey is the Australian Agricultural and Grazing Industries Survey (AAGIS), which covers farm establishments with an estimated value of agricultural operations (EVAO) of \$A22,500 or more in the last agricultural census that are classified to one of the broadacre industries – that is, cereal crop production, beef cattle production, and sheep and wool production. For the last two years, around 1650 farms have been included in the AAGIS sample, which is stratified by geographic area, industry, and EVAO. The sample farms are located throughout Australia with a non-uniform density. The latitude and longitude of the sample farms (defined in terms of the location of the farm “gate”) is recorded as a regular part of the collection. This knowledge of the location of the surveyed farms enables the spatial smoothing techniques described in this paper to be used.

Traditionally, AAGIS estimates have been presented only as tables of numbers showing averages for all Australia, each state, and industries within states. However, the concern of rural industry and government about the combined impact of drought in some areas of Australia and the decline in certain commodity prices has highlighted the need for timely and detailed information on regional trends in farm performance.

In particular, there has been a perceived need for information which portrays the spatial distribution of farm performance, reflecting actual variability in climate and production across Australia.

A highly effective way of presenting information on a spatial basis is to map the regional variation in economic performance of the surveyed farms. We use a nonparametric regression method to spatially smooth the farm level survey data, which is then presented in the form of a map. Recent improvement in computing power and the availability of high quality and affordable GIS packages have made this form of presentation a practical alternative to the traditional tabular method of presenting survey results.

Maps have been found to be a successful form of exposition for a number of reasons. First, estimates presented in a map are easily interpreted; when presented with too many tables it is very easy for a client to overlook local variations or be “swamped” by numbers. Next, maps make it easy for a client to relate the geographic variation in one variable with that of another. Finally, a colour map has great visual impact.

This demand for information on a spatial basis has resulted in an increased emphasis on small area estimates. One method of small area estimation (which originated naturally from smoothing survey data for presentation in maps) is to spatially smooth the sample weights. This reduces the variability of the small area estimates.

In Section 2, we examine a method of integrating geographical location into ABARE’s survey weighting methods in order to make our small area estimates less variable. It is applied to sub-regional estimation within two Agricultural Regions in Section 3. In Section 4, we describe how kernel regression techniques can be used to produce

<sup>1</sup> Ann Cowling, CSIRO Division of Fisheries, GPO Box 1538, Hobart TAS 7001, Australia and Australian Bureau of Agricultural and Resource Economics; Ray Chambers, Department of Social Statistics, University of Southampton, Highfield, Southampton SO17 1BJ, United Kingdom; Ray Lindsay and Bhamathy Parameswaran, Australian Bureau of Agricultural and Resource Economics, GPO Box 1563, Canberra ACT 2601, Australia.

maps which give a good indication of the local geographic variation of a surveyed variable. Two methods of mapping the smoothed data are discussed, both of which use ARC/INFO, a GIS software package. The results of a simulation study comparing various kernel regression methodologies for use in ABARE's maps are summarised in the Appendix.

## 2. SMALL AREA ESTIMATION BY SPATIALLY SMOOTHING SAMPLE WEIGHTS

The standard method used to compute sample weights at ABARE is described in Bardsley and Chambers (1984). It rests on the assumption that at some appropriate level of aggregation (say, Agricultural Region) the variable  $Y$  follows a linear model of the form

$$Y = X\beta + V \quad (2.1)$$

where  $Y$  is the  $N$ -vector of values of  $Y$  at this level of aggregation,  $X$  is a  $N \times p$  matrix of values of a set of  $p$  benchmark variables,  $\beta$  is an unknown  $p$ -vector of regression coefficients and  $V$  is a  $N$ -vector of errors satisfying  $E(V) = 0$  and  $\text{var}(V) = \sigma^2\Omega$ , where  $\sigma$  is an unknown scale parameter and  $\Omega$  is a known  $N \times N$  diagonal matrix having as its elements the measure of size of each farm, EVAO, introduced in the previous section.

Since this model is a multipurpose model, with the same set of benchmark variables used for each survey variable, the column dimension,  $p$ , of  $X$  is usually large. Typically,  $X$  consists of between 3 and 7 variables related to the main agricultural commodities produced by farms in the region together with dummy variables indicating industry strata within the region. Best linear unbiased estimation of the population total of a survey variable on the basis of such an overspecified model typically results in weights that are highly variable and often negative.

As discussed in Bardsley and Chambers (1984), negative weights are highly undesirable in a multi-purpose survey like AAGIS. In particular, such weights can lead to negative estimates of intrinsically positive quantities. This problem has been pointed out in the literature a number of times (see for example, Deville and Särndal 1992; Bankier, Rathwell and Majkowski 1992; and Fuller, Loughin and Baker 1994). The method used at ABARE to control for strictly positive sample weights is based on the ridge-type modification to the best linear unbiased weights suggested by Bardsley and Chambers (1984).

Given a sample of size  $n$  from a particular region, the ridge weighting approach determines the sample weight vector  $w$  by minimising the mean squared error criterion

$$Q = \lambda^{-1} B^T C B + (w - 1)^T \omega (w - 1). \quad (2.2)$$

Here  $B = T - x^T w$  is a  $p$ -vector of benchmark biases, corresponding to the differences between the (known)

population totals  $T$  of the  $p$  benchmark variables making up  $X$  and the corresponding survey estimates  $x^T w$  of these totals,  $C$  is a  $p \times p$  diagonal matrix of non-negative relative "costs" associated with these biases,  $\omega$  is the sample component of  $\Omega$ ,  $x$  is the sample component of  $X$ ,  $1$  is a  $n$ -vector of ones and  $\lambda$  is a scaling constant which is chosen by the survey analyst. The value of  $w$  minimising  $Q$  is

$$w = 1 + \omega^{-1} x (\lambda C^{-1} + x^T \omega^{-1} x)^{-1} (T - x^T 1). \quad (2.3)$$

The scale constant  $\lambda$  is called the ridge parameter associated with these weights. As  $\lambda$  increases from zero, the sample weights in  $w$  move away from their best linear unbiased values under the model (2.1) (namely, their values at  $\lambda = 0$ ) and become less and less variable. That is, as  $\lambda$  increases, the variances of the survey estimates based on these weights decrease. On the other hand, as  $\lambda$  increases, these estimates become more biased under (2.1), so the components of  $B$  move away from their zero values at  $\lambda = 0$  (where the sample weights define unbiased estimates under (2.1)). These components become larger and larger (in absolute terms) as  $\lambda$  increases.

The survey analyst makes a tradeoff between these two competing sources of "error" by choosing the smallest value of  $\lambda$  such that the sample weights in  $w$  stabilise at strictly positive values as close as possible to their best linear unbiased values under (2.1). This ensures that the components of  $B$  are as small as possible subject to this stability requirement. At ABARE, the value of  $\lambda$  is chosen so that the sample weights are at least unity.

Recent small area estimation research in ABARE has focussed on a method of modifying this ridge weighting procedure to create sample weights that are less spatially variable. We achieve this by modifying the mean squared error criterion  $Q$  in (2.2) to include a constraint on spatial variability, while continuing to regard the elements of the variable  $Y$  as being independent.

Let  $K$  be an  $n \times n$  matrix reflecting Euclidean distance between sample farms, such that  $K$  is symmetric and non-negative,  $K_{ii} = 1$  for all  $i$  and  $K_{ij} \downarrow 0$  as the distance between farm  $i$  and farm  $j$  increases. Put  $u = w - 1$ . The aim is then to choose  $u$  so that when  $K_{ij}$  is large, the difference between  $u_i$  and  $u_j$  is small. That is, we seek to minimise a quantity of the form

$$\sum_{i \in S} \sum_{j \in S} K_{ij} (u_i - u_j)^2 = 2(u^{(2)})^T K 1 - 2u^T K u \quad (2.4)$$

where  $(u^{(2)})_i = (u_i)^2$ . An appropriate modification to the mean squared error criterion (2.2) leads to minimisation of

$$Q^* = \lambda^{-1} B^T C B + u^T \omega u + (u^{(2)})^T K 1 - u^T K u.$$

Minimising with respect to  $u$  leads to

$$u = \eta^{-1} x (\lambda C^{-1} + x^T \eta^{-1} x)^{-1} (T - x^T 1)$$



provided  $\eta^{-1}$  exists, where

$$\eta = \text{diag}(\mathbf{K}\mathbf{1}) - \mathbf{K} + \omega. \quad (2.5)$$

Clearly, then,

$$\mathbf{w} = \mathbf{1} + \eta^{-1} \mathbf{x} (\lambda \mathbf{C}^{-1} + \mathbf{x}^T \eta^{-1} \mathbf{x})^{-1} (\mathbf{T} - \mathbf{x}^T \mathbf{1}). \quad (2.6)$$

It can be seen that the modified mean squared error criterion  $Q^*$  equally weights the spatial smoothness criterion given in (2.4), and the term corresponding to the variance of the prediction error of the sample estimates,  $\mathbf{u}^T \omega \mathbf{u}$ . As the scale of  $\mathbf{K}$  was arbitrarily specified, the comparative weighting of the two criteria must be modified by “scaling up” the spatial matrix  $\{\text{diag}(\mathbf{K}\mathbf{1}) - \mathbf{K}\}$  by a factor  $\phi$  in order to make it comparable in size with the heteroscedasticity matrix  $\omega$ , and by adding a parameter  $\alpha$ ,  $0 \leq \alpha \leq 1$ , to the expression for  $\eta$  in equation (2.5), so that

$$\eta = (1 - \alpha) \phi \{\text{diag}(\mathbf{K}\mathbf{1}) - \mathbf{K}\} + \alpha \omega.$$

These spatially smoothed sample weights can be derived in a second way, providing deeper insight into how they should be interpreted. This follows from noting that

$$\eta = \begin{bmatrix} \sigma_1^2 + \sum_{m \neq 1} K_{1m} & -K_{12} & \dots & -K_{1n} \\ -K_{21} & \sigma_2^2 + \sum_{m \neq 2} K_{2m} & \dots & -K_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ -K_{n1} & -K_{n2} & \dots & \sigma_n^2 + \sum_{m \neq n} K_{nm} \end{bmatrix}$$

can be expressed as  $\eta = \mathbf{S} \mathbf{R} \mathbf{S}$ , where  $\mathbf{S}$  is a diagonal matrix with  $S_{ii} = (\sigma_i^2 + \sum_{m \neq i} K_{im})^{1/2}$ , and  $\mathbf{R}$  is a correlation matrix with

$$R_{ij} = \begin{cases} 1 & \text{if } i = j \\ -K_{ij} \left\{ \left( \sigma_i^2 + \sum_{m \neq i} K_{im} \right) \left( \sigma_j^2 + \sum_{m \neq j} K_{jm} \right) \right\}^{-1/2} & \text{if } i \neq j. \end{cases}$$

Thus the spatially smoothed sample weights can alternatively be derived as ridge-type regression weights based on the assumption that the variable  $\mathbf{Y}$  follows a linear model of the form (2.1), with  $\mathbf{V}$  redefined as satisfying  $E(\mathbf{V}) = \mathbf{0}$ ,  $\text{var}(\mathbf{Y}_i) = \sigma_i^2 + \sum_{m \neq i} K_{im}$ , and  $\text{cov}(\mathbf{Y}_i, \mathbf{Y}_j) = -K_{ij}$  for  $i \neq j$ . The usual ridge weighting procedure then leads directly to (2.6) with  $\eta$  defined by (2.5). Note that under this implied model neighbouring farms are negatively correlated.

This second method of derivation shows clearly that the introduction of spatial smoothness for the survey weights is at odds with standard concepts of statistical efficiency as far as estimation at the aggregate level is concerned. Since the

spatial correlation between neighbouring farms will typically be positive, efficient survey estimation at the aggregate level will involve weighting based on (2.3) with  $\omega$  replaced by a non-diagonal variance/covariance matrix reflecting this positive spatial correlation. These are not the weights that result when one imposes as spatial similarity constraint. Consequently, one could expect that such “large area efficient” weights would tend to be more dissimilar for neighbouring farms than they would be for farms that are far apart. That is, there is a price to pay in weighting – if less variable aggregate level estimates are required, then this tends to lead to more variable small area estimates. Conversely, if (2.6) is adopted as the method of weighting because of its desirable small area properties, then it can be expected that aggregate level estimates obtained by summing these small area estimates will be less efficient.

The spatially smooth sample weights (2.6) have been implemented using

$$K_{ij} = \exp(-d \|z_i - z_j\|), \quad (2.7)$$

where  $\|z_i - z_j\|$  is the distance between farm  $i$  and farm  $j$  and  $d$  is a constant controlling the radius of circle around the  $i$ -th farm within which spatial smoothing is applied. The smaller the value of  $d$ , the larger the radius of spatial smoothing. At present, the “scaling up” constant  $\phi$  is computed as the ratio of the determinants of the  $\mathbf{K}$  and  $\omega$  matrices, raised to the power  $n^{-2}$ . An empirical evaluation of this method is described in the following Section.

### 3. AN APPLICATION OF SPATIALLY SMOOTHED SAMPLE WEIGHTING

Initial results from an evaluation of the first method of spatially smoothed ridge weighting described in the previous section are set out in Tables 1 to 3. These results are for two Agricultural Regions. The first, Region A, is in New South Wales. In spatial terms, this region is relatively homogeneous, being located in the southwestern corner of the state. The principal agricultural activities are wheat and rice production and wool and lamb production. The second, Region B, is in Western Australia. This region is more spatially heterogeneous, ranging from established cropping and wool production farms in the central west of the state to much larger livestock and cropping farms on marginal farming land in the south east of the state. The principal agricultural activities are wheat and legumes production and wool production.

Six variations of the spatially smoothed ridge weights (2.6) with  $\mathbf{K}$  given by (2.7) were used in the evaluation, defined by values of  $d = 0.05$  (weak spatial effects) and  $d = 0.005$  (strong spatial effects), and values of  $\alpha = 0.9$  (most emphasis on the standard ridge weights),  $\alpha = 0.5$  (equal emphasis on standard ridge weights and spatially smooth weights) and  $\alpha = 0.1$  (most emphasis on spatially smooth weights).

**Table 1**

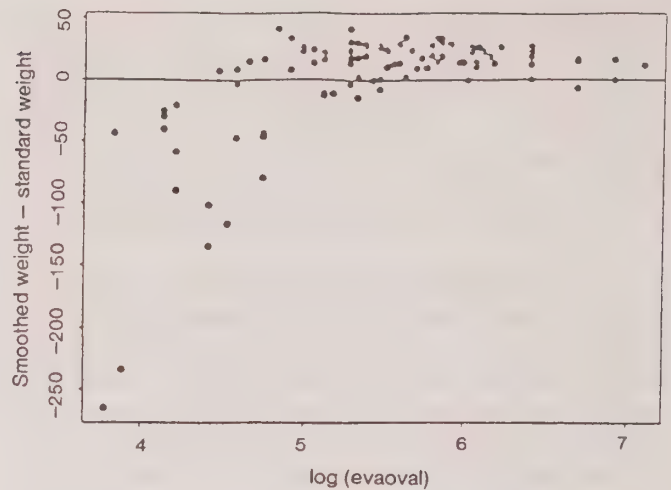
Values (in relative percentage terms) of the biases associated with estimation of the benchmark variables corresponding to the principal agricultural commodities produced in Region A (sample size  $n = 101$  farms) and Region B (sample size  $n = 85$  farms) using the standard ridge weights (2.3) and the spatially smooth ridge weights (2.6)

		Wheat	Sheep	Rice
<b>Region A</b>				
Standard ridge weights		-0.50	5.0	13.0
Spatially smoothed ridge weights				
$d = 0.05$	$\alpha = 0.9$	-0.50	4.6	11.9
	$\alpha = 0.5$	-0.46	4.7	12.4
	$\alpha = 0.1$	0.07	6.2	17.4
$d = 0.005$	$\alpha = 0.9$	-0.40	4.9	12.7
	$\alpha = 0.5$	0.80	8.9	28.0
	$\alpha = 0.1$	9.20	25.0	60.0
		Wheat	Sheep	Legumes
<b>Region B</b>				
Standard ridge weights		0.43	-1.25	1.49
Spatially smoothed ridge weights				
$d = 0.05$	$\alpha = 0.9$	0.42	-1.16	1.37
	$\alpha = 0.5$	0.44	-1.14	1.40
	$\alpha = 0.1$	0.69	-1.25	2.53
$d = 0.005$	$\alpha = 0.9$	0.50	-1.20	1.68
	$\alpha = 0.5$	1.51	1.14	9.73
	$\alpha = 0.1$	26.57	19.61	45.46

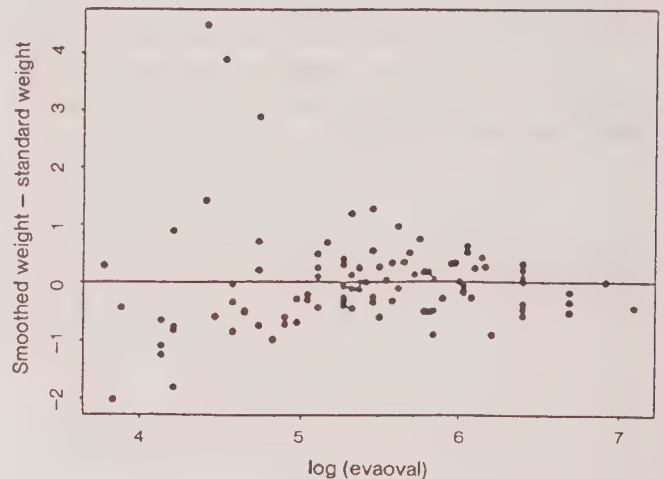
Table 1 shows the relative biases associated with estimation of the population totals of the main commodity related benchmarks for each region under these different weighting systems, as well as the corresponding biases associated with the standard ridge weights. The increase in these biases as the amount of spatial smoothing in the weights is increased is evident. Since these production benchmarks are positively correlated with most of the economic variables measured in the survey, these benchmark biases can be expected to be translated into a corresponding upward bias in survey estimates based on these weights.

Figures 1 to 4 show the difference between the smoothed weights and the standard ridge weights for the two "extreme" combinations of  $\alpha$  and  $d$  in both regions changes as the size (measured in terms of the logarithm of the estimated value of agricultural operations, or  $\log(\text{EVAO})$ ) of the sample farms changes.

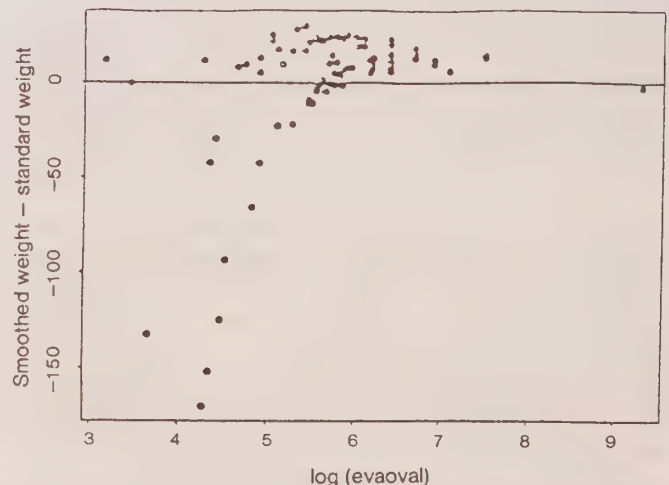
Observe that for relatively strong spatial smoothing (Figures 1 and 3), the effect of smoothing is to increase the weights of most of the larger sample farms, while dramatically decreasing the weights of a small number of smaller sample farms. Weak spatial smoothing (Figures 2 and 4) changes the weights much less, and there is little relationship between the size of the farm and the direction of weight change. Consequently, an upward shift in survey estimates for these regions could be expected with the introduction of



**Figure 1.** Difference between smoothed weight with  $\alpha = 0.1$  and  $d = 0.005$  and standard ridge weight, Region A



**Figure 2.** Difference between smoothed weight with  $\alpha = 0.9$  and  $d = 0.05$  and standard ridge weight, Region A



**Figure 3.** Difference between smoothed weight with  $\alpha = 0.1$  and  $d = 0.005$  and standard ridge weight, Region B



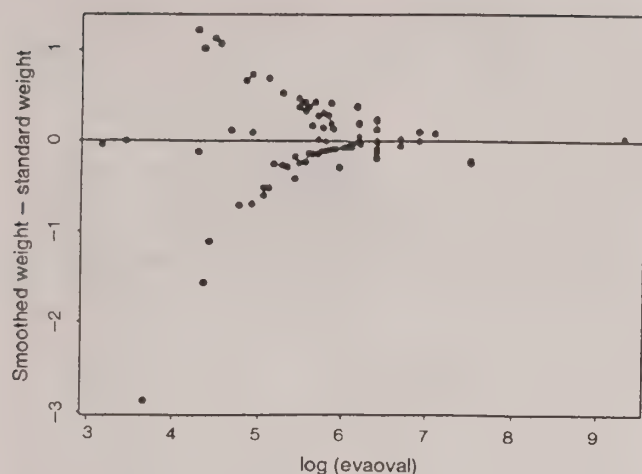


Figure 4. Difference between smoothed weight with  $\alpha = 0.9$  and  $d = 0.05$  and standard ridge weight, Region B

strongly spatially smoothed sample weights. Given the increased positive biases indicated in Table 1, this upward shift would be expected to be essentially due to the introduction of a positive bias in these estimates.

Is this increased bias compensated for by a lower standard error? To evaluate this question, survey estimates and estimated standard errors were computed for a key financial variable, total cash costs. These estimates are set out in Table 2 (Region A) and Table 3 (Region B). Estimates are provided both for each region and for small areas within each region, denoted SR- $i$  in the table, with the index  $i$  ranging between 1 and 6 for Region A and between 1 and 7 for Region B.

Table 2

Estimates (with corresponding estimated standard errors in parentheses) of the average value of  $Y$  = total cash costs in subregions SR-1 to SR-6, making up Region A (sample size  $n = 101$  farms), using the standard ridge weights (2.3) and the spatially smooth ridge weights (2.6)

	Standard weights	Spatially smoothed ridge weights					
		$d = 0.05$			$d = 0.005$		
		$\alpha = 0.9$	$\alpha = 0.5$	$\alpha = 0.1$	$\alpha = 0.9$	$\alpha = 0.5$	$\alpha = 0.1$
SR-1	100,618 (24,551)	100,453 (24,511)	101,297 (23,906)	107,263 (20,487)	102,059 (23,474)	112,635 (18,923)	135,419 (18,011)
SR-2	115,320 (26,754)	115,417 (26,661)	116,002 (26,448)	120,362 (25,637)	116,917 (26,423)	126,165 (25,990)	153,707 (27,975)
SR-3	167,524 (28,479)	167,453 (28,467)	167,486 (28,473)	168,257 (28,426)	167,709 (28,175)	170,781 (26,471)	187,683 (24,211)
SR-4	182,940 (106,471)	180,317 (105,485)	177,838 (101,012)	163,556 (74,418)	176,257 (97,823)	174,077 (69,109)	192,296 (43,651)
SR-5	132,050 (25,089)	132,083 (25,096)	132,389 (25,154)	134,786 (25,475)	132,490 (25,173)	136,369 (24,410)	151,046 (23,110)
SR-6	132,493 (44,385)	132,184 (44,546)	133,204 (44,757)	141,623 (46,736)	133,763 (45,078)	147,652 (46,953)	192,781 (53,105)
Region A	134,114 (15,691)	133,807 (15,655)	134,141 (15,426)	137,080 (13,845)	134,506 (15,199)	142,040 (13,494)	166,432 (12,815)

Table 3

Estimates (with corresponding estimated standard errors in parentheses) of the average value of  $Y$  = total cash costs in subregions SR-1 to SR-7, making up Region B (sample size  $n = 85$  farms), using the standard ridge weights (2.3) and the spatially smooth ridge weights (2.6)

	Standard weights	Spatially smoothed weights					
		$d = 0.05$			$d = 0.005$		
		$\alpha = 0.9$	$\alpha = 0.5$	$\alpha = 0.1$	$\alpha = 0.9$	$\alpha = 0.5$	$\alpha = 0.1$
SR-1	183,194 (64,851)	183,262 (64,325)	183,528 (64,051)	186,151 (64,967)	184,287 (64,132)	195,138 (69,859)	257,652 (59,518)
SR-2	261,952 (70,989)	261,487 (70,601)	261,119 (70,502)	261,182 (73,131)	261,938 (70,723)	276,912 (79,751)	331,805 (67,356)
SR-3	113,499 (30,304)	113,441 (30,289)	113,742 (30,255)	116,847 (30,731)	114,631 (30,377)	125,525 (31,507)	157,007 (32,500)
SR-4	242,220 (26,160)	242,182 (25,671)	242,208 (26,159)	242,221 (26,160)	242,163 (26,154)	242,439 (24,244)	250,871 (24,836)
SR-5	134,524 (32,420)	134,970 (32,528)	135,700 (32,432)	139,122 (30,607)	134,734 (32,202)	131,448 (27,867)	148,629 (27,942)
SR-6	176,540 (60,377)	176,977 (60,703)	175,708 (59,214)	163,241 (46,361)	172,076 (55,925)	148,434 (36,218)	171,856 (39,527)
SR-7	205,287 (44,137)	205,644 (44,008)	205,433 (43,963)	202,039 (44,044)	204,519 (43,972)	194,998 (45,434)	219,959 (51,690)
Region B	176,283 (19,039)	176,342 (18,869)	176,397 (18,874)	176,822 (18,213)	176,294 (18,511)	179,998 (18,540)	216,445 (17,099)

It is seen that, in general, the answer to the question posed above is yes. The estimated standard errors of the survey estimates decrease as the degree of spatial smoothness of the weights increases (from left to right across the tables). However, as expected, the estimates themselves also increase in size, becoming more and more positively biased. Overall, the gain due to reduced standard error seems to cancel out the increase in bias, except for the heaviest spatial smoothing ( $\alpha = 0.1$ ,  $d = 0.005$ ). In this latter case the increase in bias outweighs the reduction in standard error. The choice  $\alpha = 0.1$  and  $d = 0.05$  seems a good compromise, leading to reasonable (but not spectacular) bias-variance tradeoffs in Region A, and little change in the estimates in Region B.

#### 4. ESTIMATION AND MAPPING OF LOCAL AVERAGES

A survey data map is a two-dimensional surface which estimates the spatial mean function of the survey variable in the population. In practice, such a map is obtained by applying a nonparametric regression technique to the weighted unit record data obtained in the survey.

At ABARE, we use kernel regression (a nonparametric technique) to produce maps which show the spatial variation of the estimated spatial mean function surfaces of key survey variables. These surfaces are obtained by replacing the observed sample values of these variables by locally weighted averages. In addition, for each local average map, a

corresponding map is produced which shows an estimate of the local variability of the variable of interest. We give below a brief outline of the technique: for clarity of exposition we deal only with the univariate case. See Ruppert and Wand (1994), Wand and Jones (1995, p140), and the references therein, for discussion of the multivariate case.

We assume that the finite population is generated as an iid sample  $\{(Z_i, Y_i), i = 1, \dots, N\}$  from a super population where  $Y_i$  is the value of a response variable  $Y$  observed at location  $Z_i$ . We suppose that the observations follow the model

$$Y_i = m(Z_i) + \epsilon_i, \quad i = 1, \dots, N$$

where  $m(z) = E(Y|Z = z)$  is the conditional mean of  $Y$  given  $Z$ , and the  $\epsilon_i$  are independent random variables with zero mean and variance  $\sigma^2(z)$ . Suppose that the error terms  $\epsilon_i$  are independent of the process by which the sample is selected, so that the sample values  $\{(Z_i, Y_i), i = 1, \dots, n\}$  follow the same model, and write  $f$  for the density of  $Z_1, \dots, Z_n$ .

A natural choice for the local average at any point  $z$  is then the mean of the values of the response variable for those observations with locations close to  $z$ , since observations from points far away will tend to have very different mean values. The local average is defined as a weighted mean

$$\hat{m}(z) = n^{-1} \sum_{i=1}^n W_i(z) Y_i$$

where the weights  $\{W_i(z)\}$  depend on the locations  $\{Z_i\}$  of the sample observations, and  $\hat{m}(z)$  estimates  $m(z)$ .

The weights are constructed using a function  $K$  known as the kernel, which is continuous, bounded, symmetric and integrates to one. Various weight sequences have been proposed: the traditional Nadaraya-Watson weights (Nadaraya 1964 and Watson 1964) are

$$W_i(z) = h^{-1} K\{(z - Z_i)/h\} \left/ \left[ (nh)^{-1} \sum_{j=1}^n K\{(z - Z_j)/h\} \right] \right.,$$

where  $h$  is a scale factor known as the bandwidth. The kernel function  $K$  gives an observation close to  $z$  relatively more influence on the local average at this location than it gives to an observation further from  $z$ .

Where observations are sparse, a fixed-bandwidth window may contain few points and the corresponding estimator may therefore have a very high variance. This may be avoided by using the  $k$ -nearest-neighbour method in which a different bandwidth is used at each estimation point  $z$ . The bandwidth at  $z$  is the distance to the  $k$ -th nearest neighbour of  $z$ , so that there are always exactly  $k$  points in the bandwidth window. Let  $h_k$  be the distance between  $z$  and its  $k$ -th nearest neighbour. The  $k$ -nearest-neighbour Nadaraya-Watson weights are

$$W_{ih_k}(z) = h_k^{-1} K\{(z - Z_i)/h_k\} \left/ \left[ (nh_k)^{-1} \sum_{j=1}^n K\{(z - Z_j)/h_k\} \right] \right.$$

We show in Table 4 the asymptotic mean squared error (MSE) properties of the usual (fixed-bandwidth) and  $k$ -nearest-neighbour estimators as given in Härdle (1990, p. 46).

**Table 4**

Asymptotic bias and variance of Nadaraya-Watson estimators;  
 $c_K = \int K^2(u)du$ ,  $d_K = \int u^2 K(u)du$

	Fixed-bandwidth	$k$ -nearest-neighbour
Bias	$h^2 \frac{(m''f + 2m'f')(x)}{2f(x)} d_K$	$\left(\frac{k}{n}\right)^2 \frac{(m''f + 2m'f')(x)}{8f^3(x)} d_K$
Variance	$\frac{\sigma^2(x)}{nhf(x)} c_K$	$\frac{2\sigma^2(x)}{k} c_K$

Clearly, the bias of the estimated regression function can be reduced by using a smaller bandwidth  $h$  (number of nearest-neighbours  $k$ ), but this leads to a noisy estimate  $\hat{m}$  with local detail masking global features of the curve ( $\hat{m}$  has high variance). If  $h(k)$  is large,  $\hat{m}$  is smoother but the global features are dampened ( $\hat{m}$  has high bias and low variance). The bias, then, can only be reduced at the expense of variance and vice versa, with the bandwidth  $h$  determining the ratio of (squared) bias to variance.

In reality, the survey design and the spatial distribution of a survey variable  $Y$  will not be independent, so simple local averages for  $Y$  derived from the sample data will be misleading as estimates of the local population means of this variable. To overcome this problem the kernel weights are multiplied by the survey weights to get the final smoothing weights used for calculating the local average. This is equivalent to estimating the local population mean  $m(z)$  of  $Y$  under the assumption that it is locally linear in the same benchmark variables as those used to model the overall population mean of  $Y$ .

A wide array of alternative kernel smoothing procedures have been discussed in the literature. As well as various sequences of smoothing weights  $\{W_i\}$ , there are different types of bandwidths, and several automatic bandwidth selection methods. A simulation study was therefore conducted to determine the most appropriate kernel methodology for use in ABARE's maps. This is described in the Appendix.

Uncertainty about the estimate of the spatial mean derived via kernel-based spatial smoothing can be represented by mapping the local variability of the variable of interest. Areas of high local variability correspond to areas where the map of the mean function is less precise and vice versa for areas of low local variability.

The usual method of determining confidence regions for a kernel curve estimate is the bootstrap; see Härdle (1990), Hall (1992), and references therein. However, for computational efficiency, we use the expectiles (Newey and Powell 1987) of the spatial distribution of  $Y$  to describe this



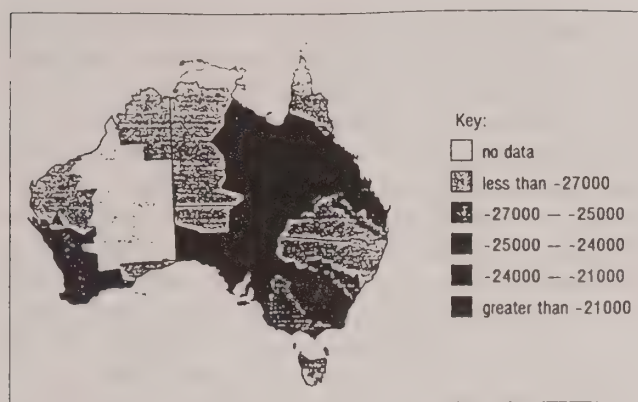


Figure 5. Polygon map of farm business profit in 1991-1992, all broadacre farm (\$)

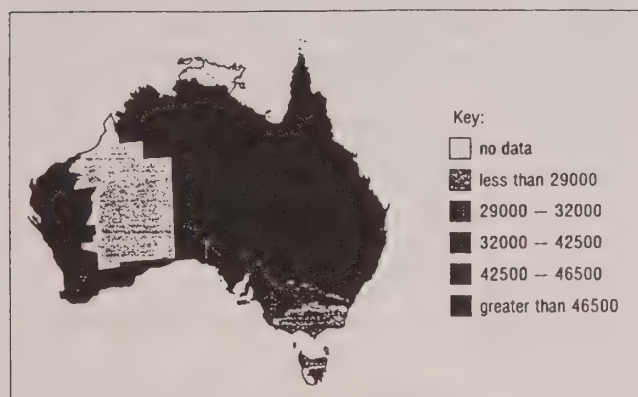


Figure 6. Polygon map of interexpectile range of farm business profit in 1991-1992, all broadacre farms (\$)

local variability. An expectile bears the same relationship to the mean as the corresponding quantile does to the median. In particular, the difference between the 75th and 25th expectiles of a distribution is a measure of the spread of the distribution in the same way as the interquartile range is a measure of this spread. The smoothing program contains a module for non-parametric  $M$ -quantile regression (Breckling and Chambers 1988) which is used to fit a smooth surface to the expectiles of the  $Y$ -distribution at any location. The difference between the smoothed 75th and 25th expectile surfaces (the smooth expectile analogue of the interquartile range) is then mapped to show areas of high and low variability in the data.

Not surprisingly, this smooth interexpectile range tends to be highest in areas where the farms are sparsely located and the farm-to-farm variability in  $Y$  is therefore highest. The interexpectile range map corresponding to Figure 5 is shown in Figure 6. Note that these smoothed interexpectile range maps provide similar information to confidence bands at any particular point on the map. However, they do not have the same repeated sampling interpretation as confidence intervals, and hence should be treated as guides to, rather than measures of, the uncertainty associated with a particular map.

For confidentiality reasons, care must be taken when mapping the smoothed data for publication to ensure that the locations of the surveyed farms are not revealed. Another requirement is output quality compatible with desktop publication packages. Two procedures for generating the final maps that satisfy these requirements have been developed using ARC/INFO.

In the first method, a Thiessen polygon is constructed around each farm. The polygon defines the area closer to that farm than to any other farm. The farm location is not in the centre of its polygon, and the polygon shape does not resemble the shape of the farm, so the polygons conceal the locations of the survey farms, as shown in Figure 7. The whole of each polygon is coloured according to the smoothed value of  $Y$  at the farm location in that polygon. Usually ten colours are used in each map and the estimated population deciles of the smoothed data are used as boundaries for the colour area. The maps shown in this paper are black-and-white analogues of these colour maps.

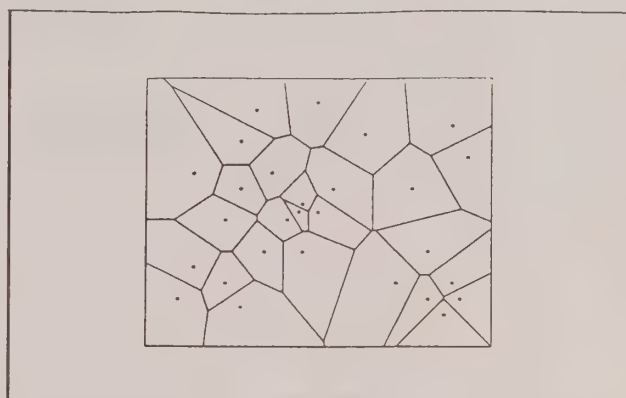


Figure 7. Thiessen polygons constructed around selected ABARE survey farms. Farm location is shown as a small square within each polygon

In the second method, smoothed values on a dense rectangular grid are used in place of smoothed values at the farm locations, and a further minor interpolation of the data is carried out in ARC/INFO. A continuous 3-dimensional surface which passes through the smoothed values at the grid points is built in two steps. As a first approximation, a faceted surface of triangles obtained by Delauney triangulation is constructed, and then a bivariate fifth degree polynomial is fitted within each triangle using Akima's algorithm (Akima 1978). The resulting continuous surface is then contoured using the estimated population deciles. Figure 8 is an example.

In this second method of presentation, the locations of the survey farms are not used in any way, thereby completely concealing the location of each survey farm. It also gives smooth contours, and the result is not as patchy as the polygon based map. Moreover, it is preferred by ABARE's graphics staff because it reduces the number of areas to be

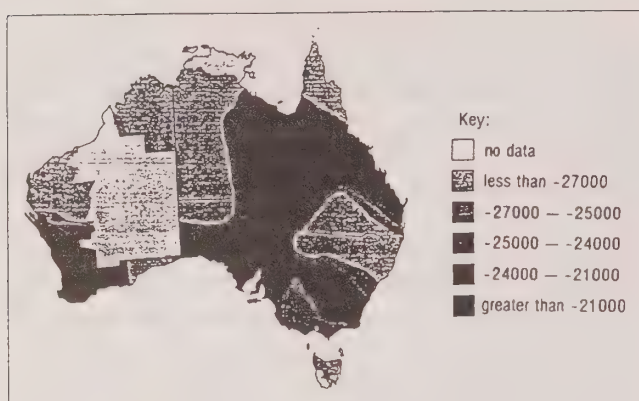


Figure 8. Contour map of farm business profit in 1991-1992, all broadacre farms (\$)

separately coloured and has lower storage requirements, enabling the maps to be more readily manipulated in desktop publishing packages. Its disadvantage is that it uses more computing time in the ARC/INFO stage.

Since the above procedures interpolate across all of Australia, including areas where there is no agricultural activity, the final stage of the map production in ARC/INFO is the “blanking out” of those areas of Australia where there are few or no farms involved in the particular broadacre industry represented by the map. As Figure 9 shows, different areas are blanked out for different industries.



Figure 9. Polygon map showing expected change in wool production, 1991-92 to 1992-93, farms with 100 or more sheep in 1991-92 (kg)

## 5. DISCUSSION

In this paper we have demonstrated that when survey data has a spatial dimension, as in the case of the AAGIS, spatial smoothness concepts may be useful to the analyst. The concept can be used to modify survey weights to ensure less variable small area survey estimates. It may also be used to smooth the data along spatial dimensions before mapping the spatial mean function.

Because we describe mapping in this paper, we have only considered smoothing along spatial dimensions. However, it is clearly possible to use the same techniques to smooth along other dimensions. Thus, if there is reason to expect the presence of strong serial correlation when the underlying population is ordered according to some variable, then one can consider applying the methods described in this paper to mapping the “change” in the survey variables relative to the change in this variable. In doing so, it should be noted that such “maps” are nothing more than nonparametric estimates of the conditional means of the survey variables given this “ordering” or “smoothing” variable. The analyst should, however, remember the “curse of dimensionality”: the effective sample size drops sharply with each additional smoothing variable used in these nonparametric techniques.

Finally, in mapping the survey data, we have used kernel-based estimation techniques. However, spline smoothing, or even parametric methods could also be used. We regard the choice of smoothing technology as somewhat subjective and purpose specific, as there are no definitive objective reasons for preferring one method over another.

## ACKNOWLEDGEMENTS

The authors would like to thank the referees for their helpful comments which have greatly improved the presentation of the paper.

## APPENDIX

In the last few years a number of optimality properties have been established for the locally-linear kernel weights (see for example Wand and Jones (1995) and references therein). We therefore compared Nadaraya-Watson (NW) and locally-linear (LL) weight sequences using fixed (FBW) and  $k$ -nearest-neighbour (NN) bandwidths with each weight sequence. For each of these combinations, we selected the bandwidth using least-squares cross-validation (CV), and an *ad hoc* method (detailed in the last paragraph of this section) aimed at reducing the speckledness of a map (SF).

Two criteria were used to evaluate the performance of each methodology. The first, MSE, is the obvious statistical criterion for assessing a biased estimator. The second criterion is more ABARE specific. As estimates are produced both in tables (by State) and in maps, the impression of the state average given by the map should be close to the tabulated value. We therefore used a weighted sum of the squared differences between the state averages of the raw and smoothed survey data (SB<sup>2</sup>). This measure was also calculated at regional rather than state level (RB<sup>2</sup>; there are between one and nine regions in each state).

Data were generated at the survey farm locations using three smooth functions with varying degrees of smoothness (measured by  $\int m''$ ) and normal mixture errors. For example,



$$m_1(z) = 6.25 \times 10^4 \times \cos\left(\frac{z_1 - 132.5}{2.25}\right) \cos\left(\frac{z_2 + 27.5}{1.75}\right),$$

where  $z_1$  and  $z_2$  are the longitude and latitude of the point  $z$ . The functions  $m_i(z)$  were scaled to have the same range as the smoothed values of a key survey variable, and the errors were scaled to have the same range as the residuals of the same variable after smoothing. Large variances were generated at locations with high residuals, and small variances at locations with low residuals. The simulation results based on the smooth function are given in Table 5.

Using MSE as the criterion for assessing methodology, the results were not consistent for the three functions  $m_i(z)$ . However, when either  $RB^2$  or  $SB^2$  was used as the performance measure, the LL estimator with  $k$ -nearest-neighbour bandwidth selected using SF outperformed the other methods by at least ten percent for each function  $m_i(z)$ , and is therefore the currently preferred methodology for producing ABARE's maps.

Table 5

Comparison of locally-linear (LL) and Nadaraya-Watson (NW) weight sequences, using fixed (FBW) and  $k$ -nearest-neighbour (NN) bandwidths selected by least-squares cross-validation (CV) and the criterion detailed below (SF). The results were obtained from 400 independent samples with mean function and normal mixture errors. The MSE values were calculated using the average over the finite population of  $(y - \hat{m}(z))^2$

		MSE $\times 10^{-7}$		RB <sup>2</sup> $\times 10^{-7}$		SB <sup>2</sup> $\times 10^{-7}$	
		CV	SF	CV	SF	CV	SF
LL	FBW	39.64	93.93	4.44	1.67	1.33	0.39
	NN	20.50	22.83	2.22	1.35	0.37	0.14
NW	FBW	41.91	52.78	3.29	1.77	0.34	0.17
	NN	21.77	22.22	3.03	2.33	0.62	0.41

The bandwidth selection method aimed at reducing the speckledness of a map (SF) is a measure of the smoothness of the map: it measures how similar the smoothed value is at any farm to that of its neighbours. Let  $p(i)$  be the survey estimate of the percentile of the smoothed variable at the  $i$ -th farm. Let  $S_i$  be the set of indices of the six farms closest to the  $i$ -th farm. In this method, the value of

$$SF(h) = (6n)^{-1} \sum_{i \in S_i} |p(i) - p(k)|$$

is calculated. It is scale-free, and decreases monotonically as the bandwidth decreases. The chosen bandwidth is the smallest bandwidth with a sufficiently small ( $< \epsilon$ ) rate of decrease of SF. The value of  $\epsilon$  was chosen subjectively following detailed examination of maps of five key variables for five values of  $\epsilon$ .

## REFERENCES

- BANKIER, M.D., RATHWELL, S., and MAJKOWSKI, M. (1992). Two step generalised least squares estimation in the 1991 Canadian Census. *Proceedings of the Workshop on Uses of Auxiliary Information in Surveys*. Statistics Sweden, Örebro, October 5-7.
- BARDSLEY, P., and CHAMBERS, R.L. (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics*, 33, 290-299.
- BRECKLING, J., and CHAMBERS, R.L. (1988).  $M$ -quantiles. *Biometrika*, 75, 761-771.
- DEVILLE, J.-C., and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- FULLER, W.A., LOUGHIN, M.M., and BAKER, H.D. (1994). Regression weighting in the presence of nonresponse with application to the 1987-1988 Nationwide Food Consumption Survey. *Survey Methodology*, 20, 75-85.
- HALL, P. (1992). *The Bootstrap and Edgeworth Expansion*. New York: Springer-Verlag.
- HÄRDLE, W. (1990). *Applied Nonparametric Regression*. Cambridge: Cambridge University Press.
- NADARAYA, E.A. (1964). On estimating regression. *Theory of Probability and its Applications*, 10, 186-190.
- NEWHEY, W.K., and POWELL, J.L. (1987). Asymmetric least squares estimation and testing. *Econometrica*, 55, 819-847.
- RUPPERT, D., and WAND, M.P. (1994). Multivariate locally weighted least squares regression. *Annals of Statistics*, 22, to appear.
- WAND, M.P., and JONES, M.C. (1995). *Kernel Smoothing*. London: Chapman and Hall.
- WATSON, G.S. (1964). Smooth regression analysis. *Sankhyā*, Series A, 26, 101-116.





## Using Data on Interruptions in Telephone Service as Coverage Adjustments

J. MICHAEL BRICK, JOSEPH WAKSBERG and SCOTT KEETER<sup>1</sup>

### ABSTRACT

Telephone surveys in the U.S. are subject to coverage bias because about 6 percent of all households do not have a telephone at any particular point in time. The bias resulting from this undercoverage can be important since those who do not have a telephone are generally poorer and have other characteristics that differ from the telephone population. Poststratification and the other usual methods of adjustment often do not fully compensate for this bias. This research examines a procedure for adjusting the survey estimates based on the observation that some households have a telephone for only part of the year, often due to economic circumstances. By collecting data on interruptions in telephone service in the past year, statistical adjustments of the estimates can be made which may reduce the bias in the estimates but which at the same time increase variances because of greater variability in weights. This paper considers a method of adjustment using data collected from a national telephone survey. Estimates of the reductions in bias and the effect on the mean square error of the estimates are computed for a variety of statistics. The results show that when the estimates from the survey are highly related to economic conditions the telephone interruption adjustment procedure can improve the mean square error of the estimates.

**KEY WORDS:** Coverage; Bias; Weighting adjustment; Telephone sampling; RDD surveys.

### 1. INTRODUCTION

Telephone surveys provide a relatively economical method of data collection compared with face-to-face interviewing. However, telephone surveys in the U.S. are subject to an important source of bias that does not affect household surveys conducted with face-to-face interviewing: at present only 94 percent of households nationally have telephone service at any given time. Moreover, for some populations such as households with young children, coverage rates are even lower.

Weighting that includes poststratification based on demographic variables known to be associated with telephone coverage is effective in mitigating some of the consequences of coverage bias in telephone surveys. Postsurvey weighting is also generally used to compensate for nonresponse and other biases. But even when effective, weighting to known demographic totals only partially solves the problem of coverage bias, undercompensating for some variables (Massey and Botman 1988) and overcompensating for others (Brick, Burke, and West 1992).

This article describes a study of an alternative method for adjusting telephone survey data to compensate for coverage bias. The method, suggested by Keeter (1995), is based on the observation that telephone subscription is a dynamic condition not just across households in the population, but also within many households over time. A sizable number of U.S. households lose and gain telephone status during a given year. Because of this phenomenon, the telephone population at a given time includes households that have recently been in the

nontelephone population. Despite considerable information on the size and characteristics of the nontelephone population, little is known about its dynamics over shorter time periods. Evidence from social workers, telephone companies, and others who deal with indigent households suggests that for many families, telephone subscription is episodic. Households may have a telephone when they can afford it, but the telephone may be turned off when times are harder, or when the bills get too large to manage, (Federal Communications Commission 1988). It is not known how many households change their telephone status and how long they stay in a particular status.

Keeter (1995) examined two household panel surveys to obtain estimates of the dynamics of telephone service subscription. Those households that changed telephone status (presence of a telephone in the household) are called 'transient' households. For data from one panel survey that collected data 12 months apart, half of the 6 percent of all households without a telephone at either time were transient. For the other panel survey in which data were collected only two months apart, one-fourth of the 6 percent of households without telephones at either point in time were transient. Since these estimates were based on observations at two points in time rather than continuous measurement, they underestimate the percent of households that are transient. Nevertheless, these results show that a substantial proportion of households without a telephone at a specific point in time is transient.

Another important condition that must be satisfied if the transient telephone households are to be useful in reducing

<sup>1</sup> J. Michael Brick and Joseph Waksberg, Westat, Inc., 1650 Research Blvd., Rockville, MD 20850, U.S.A.; Scott Keeter, Virginia Commonwealth University, Survey Research Laboratory, Richmond, VA 23284, U.S.A.

coverage bias involves the characteristics of transient households and nontelephone households. If the two groups are not similar, then the adjustments will not be effective. Using the panel data and data from several Virginia surveys, Keeter (1995) showed that the characteristics of the transient households are much more consistent with nontelephone households than telephone households.

These findings suggest the possibility that weighting adjustments that use the data from households that have telephones only sometimes during the year might be an improvement over the current practice. To evaluate this approach to adjusting the weights, questions were added to two national surveys conducted in 1993 by Westat. Both of these surveys were random digit dial (RDD) and computer assisted telephone surveys, and the data were collected in the telephone research centers of Westat.

One of the surveys is the National Household Education Survey of 1993 (NHES:93). The NHES:93 was conducted for the National Center for Education Statistics of the Department of Education in the spring of 1993 to study issues related to school readiness of young children and school safety and discipline of children in school. The other survey was the National Survey of Veterans (NSV) which was conducted in the second half of 1993 for the U.S. Department of Veterans Affairs. In this survey, adults were screened to determine if they were veterans, and the veterans were then asked about a variety of topics including their health, education, and financial status.

Below, we present estimates of the percentage of persons that experienced some interruption of telephone service, describe procedures for adjusting the survey weights using these data, and discuss the statistical implications of using the adjusted weights. The final section summarizes the findings and gives some considerations for using this technique in RDD telephone surveys.

## 2. ESTIMATES OF INTERRUPTIONS OF TELEPHONE SERVICE

Estimates of the percentage of persons with interruptions of telephone service from national surveys were needed to further examine the potential of reducing coverage biases using these data. Questions were added to the NSV and the NHES:93 for this purpose. In the NSV, about 23,000 households were screened and interviews were completed with over 5,500 eligible veterans. In the screening interview, all household members 14 years and over were enumerated and questions were asked about their characteristics and their veteran status. If a sampled adult was a veteran, then a more detailed interview was attempted. The results reported here are those asked about the adults enumerated in the screening interview which included only a few characteristics of the adults and the household.

In the NHES:93, 64,000 households were screened and nearly 30,000 interviews were conducted within those screened households. Two survey components were included:

School Readiness (SR) and School Safety and Discipline (SS&D). Approximately 11,000 parents of 3- to 7-year-olds completed interviews on SR topics and about 12,700 parents of children in grades 3 through 12 were interviewed for the SS&D component. Data on interruptions in telephone service were collected from households in which at least one SR or SS&D interview was completed.

Since the responses to the questions in the NHES:93 were only obtained for those households that completed either an SR or SS&D interview, many characteristics of the children can be analyzed, but the data do not apply to as broad a population as the NSV. The NSV applies to all adults, but only limited data were collected on most of the adults. For all households that had completed an interview (a screening interview in the NSV and a more detailed interview in the NHES:93), a member of the household was asked if the household had experienced an interruption in telephone service in the last 12 months and how long it lasted.

### Estimated Service Interruptions in the NSV and NHES:93

The estimated percentage of persons in households that had a telephone interruption of one day or more during the last 12 months varies substantially from survey to survey. Only 2.3 percent of adults had an interruption of one day or more based on the data from the NSV, while the percentage from the NHES:93 for younger children (the SR population of 3- to 7-year-olds) was 12.0 percent, and for the SS&D population of older children (grade 3 through 12) it was 9.2 percent.

Figure 1 shows estimates and 95 percent confidence intervals of the percentage of persons that had interruptions of one day or more along with estimates for those with interruptions of telephone service that lasted for at least one week and at least 4 weeks. While the percentages vary from sample to sample, the patterns of increase by length of interruption are relatively stable. The percentage with interruptions of one week or longer is less than half the percentage with any interruption, and the percentage with interruptions of 4 weeks or more is about one-fourth the percentage with any interruption.

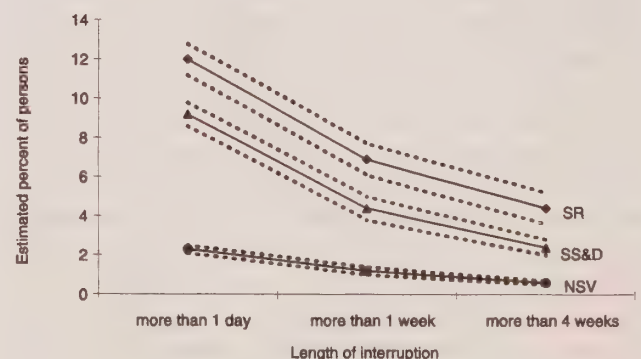


Figure 1. Estimated percentage of persons with interrupted telephone service from the three populations



The large difference in the estimates from the NSV and the NHES:93 comes from at least two important sources. The first source is that the populations were different. We would expect young children to live in households that experience more interruptions than older children and adults. Thornberry and Massey (1988) estimated that the telephone coverage rate for young children was lower than for any other age group. Thus, the difference of about 3 percent in the estimates of the percentage with an interruption between the younger (SR) and older (SS&D) children from the NHES:93 is reasonable.

The difference in the populations does not completely account for the large difference between the NSV and the NHES:93 estimates. An important reason for this difference is related to the way the questions were asked in the two surveys. The NHES:93 interview began by asking, "During the past 12 months, has your household ever been without telephone service for more than 24 hours?". In the NSV interview, respondents were asked if, "At any time during the past 12 months, has your household *not* had telephone service?". This was followed by a question that asked if the interruption was for at least 24 hours. Thus, the NSV version was a screening item followed by a more detailed question. This type of construction often depresses reports of subsequent activities, which is consistent with the lower NSV estimates.

A more important reason for the difference is probably due to the wording of the questions. With the NSV question, a 'no' response may have confused respondents because the question asks if they did *not* have telephone service. Converse and Presser (1986) discuss the problems that arise with this type of question construction. The wording for the NHES:93 is less confusing. The combination of the wording and the use of a screening item in the NSV is likely to be the main reason for the smaller estimate using the NSV questionnaire.

The difference in the estimates associated with the different ways of asking the interruption questions is evident from the estimates from two surveys conducted in Virginia by Virginia Community University. In a November 1993 survey, the items about telephone interruptions were asked using the NSV wording; in April 1994 the items were changed to the NHES:93 wording. The results from the surveys parallel the differences in the estimates between the NSV and the NHES:93. The November 1993 Virginia study estimated that 3 percent had an interruption in service in the last 12 months, while in April the estimated percentage was 9 percent. Thus, it is clear that the different ways of asking the questions heavily influenced the size of the estimates, and it suggests that the estimates from the NSV are biased downward. Some adults who did experience an interruption in telephone service during the previous 12 months probably responded incorrectly in the NSV.

### Characteristics of Persons With Service Interruptions

Estimates of the percentage of persons who had a telephone interruption are examined below by the characteristics of the person to evaluate the potential of using these data to adjust for coverage bias. We estimated the percentage of

persons in households with any interruption in service by characteristics collected in both the NSV and the NHES:93. These estimates are shown in the first part of table 1. Some differences in the distributions may be due to the different ways of asking the questions. For example, the education classification is different in the two surveys: in the NSV education is recorded for the oldest person in the household, while in the NHES:93 education is the highest for either of the parents of the child.

All subsequent analysis is restricted to NHES:93 data for two reasons. First, more data on the characteristics are available from the NHES:93 detailed SR and SS&D interviews than the NSV screening interview. Second, the telephone interruption estimate from the NSV is biased due to the wording of the item, as discussed earlier. Of course, the NHES:93 estimates apply to households with children which have higher nontelephone rates than the general population, and in that sense they do not reflect the situation for the total population.

Using the NHES:93 data, we find that the percents of persons with some interruption are relatively consistent for the SR and the SS&D populations (see table 1). The characteristics generally associated with lower economic status have the highest percentage with interruptions. For example, the percentage of children with interruptions in both the SR and SS&D populations is larger for those from households with lower household income than for those from households with higher income. Similarly, children participating in public assistance programs (WIC or free meals) have much higher rates of service interruptions than nonparticipants. However, the percentages of children in households with telephone interruptions are less variable for characteristics related to school readiness and school safety and discipline than for the socioeconomic items. Additional characteristics for both populations were examined and presented in Brick, Keeter, Waksberg and Bell (1996), but are not shown here. For most of the other substantive items, the differences in the percentage of persons with some interruption in telephone service were either not statistically significant or not large enough to be of great practical importance.

### 3. WEIGHT ADJUSTMENTS

In almost all sample surveys, the data collected from respondents are adjusted to account for nonresponse and noncoverage and to reduce the variability in the estimates by using auxiliary data from other data sources. One of the most important benefits of this type of adjustment in telephone samples is that it often reduces the bias associated with the undercoverage of persons living in households without telephones.

Kalton and Kasprzyk (1986) discuss adjustments to the base weights, classifying the adjustments into four categories: population weighting adjustments, sample weighting adjustments, raking ratio adjustments, and response probability

**Table 1**  
Estimated Percentage of Persons With Any Interruptions in Telephone Service in Last 12 Months for Three Populations

	NSV		NHES:93 SR		NHES:93 SS&D	
	Estimate	Standard error	Estimate	Standard error	Estimate	Standard error
Total	2.3	0.1	12.0	0.4	9.2	0.3
Region						
Midwest	2.3	0.2	11.0	1.0	7.3	0.7
Northeast	2.0	0.2	9.5	1.2	9.0	0.8
South	2.6	0.2	13.6	0.7	10.8	0.6
West	2.4	0.2	12.5	0.9	9.2	0.8
Race/ethnicity <sup>1</sup>						
White	2.0	0.1	9.3	0.5	7.2	0.3
Black	3.5	0.4	19.8	1.5	14.7	1.1
Hispanic	3.9	0.5	17.2	1.5	14.1	1.1
Other	2.6	0.6	11.7	2.6	9.3	1.5
Education <sup>2</sup>						
Less than high school diploma	3.2	0.2	18.4	1.8	17.4	1.6
High school graduate	2.0	0.2	15.4	0.8	11.0	0.8
Some college	2.3	0.2	11.8	0.7	8.6	0.5
Bachelor's degree	1.6	0.2	5.5	0.8	5.3	0.8
Graduate school	2.2	0.3	5.2	0.7	4.5	0.6
Household income						
\$10,000 or less			22.8	1.3	19.0	1.3
\$10,001 to \$20,000			19.9	1.4	15.7	1.1
\$20,001 to \$30,000			9.3	0.8	7.9	0.6
More than \$30,000			5.5	0.5	5.0	0.3
Women, infant and children program participant <sup>3</sup>						
Yes			18.2	1.3		
No			8.0	0.6		
Free meal at school or center <sup>4</sup>						
Yes			21.1	1.2		
No			7.6	0.5		
Birth weight						
5.5 pounds or less			12.0	1.6		
Greater than 5.5 pounds			12.0	0.4		
School control						
Public					9.4	0.4
Private					7.5	1.1
Ease of obtaining marijuana at school <sup>5</sup>						
Very or fairly easy					9.7	0.6
Hard					8.0	0.8
Nearly impossible					9.0	0.7

<sup>1</sup> Race/ethnicity is reported for the oldest member in the NSV and for the child in the NHES:93.

<sup>2</sup> Education is for the oldest household member in the NSV and the most educated parent of the child in the NHES:93.

<sup>3</sup> Estimate restricted to preschoolers.

<sup>4</sup> Estimate applies to children except preschoolers.

<sup>5</sup> Estimate applies only to children in grades 6 through 12.

Source: U.S. Department of Veterans Affairs, National Survey of Veterans, summer/fall 1993, and U.S. Department of Education, National Household Education Survey, spring 1993.



adjustments. In the NHES:93, sample weighting adjustments and raking ratio adjustments were used. Sample weighting adjustments were used to account for differential nonresponse from sampled persons. Raking ratio adjustments were then used to make the specified marginal distributions of the sample correspond to totals from the October 1992 Current Population Survey (CPS). One of the most important benefits of the type of raking ratio adjustment used in the NHES:93 is that it reduces the bias associated with the undercoverage of persons living in households without telephones because the CPS covers persons in both telephone and nontelephone households.

The data on telephone service interruptions can be used to make a response probability adjustment. Response probability adjustments are constructed by assuming that each sampled unit has a probability of responding to the survey, estimating that probability, and then using the inverse of the estimated response probability as a weighting adjustment. The Politz and Simmons (1949) method is probably the best known application of the response probability adjustment procedure, and Kalton and Kasprzyk (1986) discuss others.

To apply this type of adjustment using the telephone service interruption data, assume that living in a telephone household is a dynamic phenomenon and that a probability distribution can be associated with this status. Conceptually, a survey is conducted by sampling from this distribution and observing only those members that live in telephone households at the time of the survey. The probability of living in a telephone household (the equivalent of the response probability) must then be estimated for each respondent. The inverse of the estimated probability is the coverage adjustment. This model assumes that each person can be assigned a probability of being in a household with a telephone and that the probability is between zero and one (but not equal to zero).

The data on whether or not a household had an interruption in telephone service and the length of that interruption are the basis for this type of adjustment. Persons are divided into two categories: those in households with interruptions in service and those in households without interruptions in service. The probability is assumed to be equal to one for persons in households without interruptions and their weights are not adjusted. The weights of persons in households with at least some interruptions in the last 12 months are adjusted to account for other households that have a probability of being covered of less than one. The adjustments may vary depending on the length of time they lived in nontelephone households and on other characteristics of the household. The purpose of having different adjustments is to account for the fact that some persons are more likely to live in nontelephone households than others.

Although the weighting adjustments may reduce the undercoverage bias, introducing adjustments also typically increases the variances of the estimates. Kish (1992) discusses the reasons for unequal weights as well as the consequences from using them in a variety of situations. He advocates a common statistical approach of balancing the bias reductions against

the variance increases. If the weights reduce the bias of the estimates significantly, then it may be worthwhile accepting the variance increases. On the other hand, small reductions in bias associated with large variance increases are not recommended.

In the remainder of this section, the specific weighting adjustment procedures are described. The statistical properties of the weights developed under four alternative adjustment schemes are presented. The alternative weights are applied to the NHES:93 data and the decrease in the bias of the estimates is compared with the increase in the variance of the estimates due to the unequal weighting.

### Adjustment Schemes

The first step was to decide how to classify the length of interruption in telephone service. Various lengths of interruptions were examined to determine cut-offs that discriminated between temporary interruptions, not due to economic causes and others. It was decided to use two categories for forming adjustment cells: one week or more, and one month or more.

Within each of the length-of-service interruption categories, the children were classified into adjustment cells based on either parental education or tenure (home ownership). Race/ethnicity was used to form cells within the parental education and tenure categories. These cells were chosen because the percentage of persons with interruptions varied by these characteristics and the corresponding data were also available from the CPS. Four adjustment schemes were defined using these items:

**Scheme A1** – children in households that had a telephone service interruption of one week or more within categories defined by parental education (less than high school, high school diploma, college diploma or above) and race/ethnicity (Hispanic, black/non-Hispanic, white and other/non-Hispanic);

**Scheme A2** – children in households that had a telephone service interruption of one month or more within categories defined by parental education and race/ethnicity;

**Scheme B1** – children in households that had a telephone service interruption of one week or more within categories defined by tenure (own/other, rent) and race/ethnicity; and

**Scheme B2** – children in households that had a telephone service interruption of one month or more within categories defined by tenure and race/ethnicity.

The adjustment factors for these schemes could not be obtained directly from the NHES:93 data because no data were collected from households without telephones. Instead, the adjustments were developed using both CPS and NHES:93 data and then applied to the NHES:93 weights.

To motivate the adjustment of the weights, consider partitioning the universe of persons into four components:  $t_1$  is the number of persons in *telephone* households with *no telephone interruptions* in the past year;  $t_2$  is the number of persons in *telephone* households with *some telephone interruptions* in the past year;  $t_3$  is the number of persons in *nontelephone* households with *no telephone interruptions* in

the past year (*i.e.*, persons who lived in nontelephone households throughout the entire year); and  $t_4$  is the number of persons in *nontelephone* households with *some telephone interruptions* in the past year. As noted above, the response probability model assumes  $t_3 = 0$ .

Using the CPS it is possible to estimate  $t_1 + t_2$  and  $t_4$  (assuming  $t_3 = 0$ ); designate these estimates as  $\hat{t}_1 + \hat{t}_2$  and  $\hat{t}_4$ , respectively. From the NHES:93,  $t_1$  and  $t_2$  can be estimated separately; call these estimates  $t_1^*$  and  $t_2^*$ , respectively. The bias in the NHES:93 estimates arises because they are from a telephone survey and do not include persons in nontelephone households ( $t_4$ ).

A weight adjustment of  $A = 1 + t_4/t_2$  would result in unbiased estimates of totals; however, this adjustment involves unknown, population quantities that must be estimated. Since  $t_2$  can only be estimated from the NHES:93 and  $t_4$  can only be estimated from the CPS (assuming  $t_3 = 0$ ), the adjustment is expressed in ratios to reduce the bias due to estimating the totals from different surveys. The revised weight is

$$w'_i = w_i \left( 1 + \delta_i \frac{\frac{\hat{t}_4}{\hat{t}_1 + \hat{t}_2}}{\frac{t_1^*}{t_2^*}} \right), \quad (1)$$

where  $w_i$  is the NHES:93 weight adjusted for nonresponse of sampled persons but not yet raked to October 1992 CPS totals,  $\delta_i = 1$  if the person lives in a household that had an interruption of telephone service in the last year and is zero otherwise. The quantity in parenthesis in (1) is an estimate of  $A$ , the weight adjustment.

Revised weights were computed separately for the SR and SS&D components. Rather than the overall adjustment as given in (1), the weight adjustments were computed within the cells defined for each of the four weighting schemes (A1, A2, B1, and B2). Table 2 shows the resulting adjustment factors for the SR and SS&D components. The adjustments in the first column are those for schemes A1 and B1. The second column contains the adjustment factors for schemes A2 and B2. The adjustment factors for the schemes based on the one month or more interruptions are greater than those based on the one week or more because the denominator of the ratio is, by definition, smaller for this classification (see Figure 1 for estimates of the percentage of persons with interruptions for each scheme).

The last weighting step rakes the four alternative weights to the same October 1992 CPS totals used in raking the standard NHES:93 person-level weights. The result of this process is the standard NHES:93 weight and four alternative weights based on different adjustment schemes. All five of the weights conform to the same marginal totals. The only difference in the weights is the adjustment for the telephone

**Table 2**  
Weighting Cell Adjustments Factors, Based on Length of Interruption of Telephone Service

Factor	SR		SS&D	
	Length of service interruption			
	One week or more	One month or more	One week or more	One month or more
Cells defined by parental education and race/ethnicity (Schemes A1 and A2)				
Less than high school; Hispanic	5.75	16.35	4.89	8.52
Less than high school; black, non-Hispanic	5.10	6.72	4.26	5.95
Less than high school; white and other, non-Hispanic	4.98	5.37	3.81	4.86
High school diploma; Hispanic	2.31	2.76	2.67	4.51
High school diploma; black, non-Hispanic	2.65	3.73	3.06	4.71
High school diploma; white and other, non-Hispanic	2.16	2.79	2.18	3.09
College degree or more; Hispanic	1.34	2.33	1.96	8.22
College degree or more; black, non-Hispanic	1.77	2.64	1.35	8.83
College degree or more; white and other, non-Hispanic	1.58	2.09	1.91	3.48
Cells defined by tenure and race/ethnicity (Schemes B1 and B2)				
Renter; Hispanic	3.74	5.15	3.58	6.08
Renter; black, non-Hispanic	3.23	4.54	3.38	4.95
Renter; white and other, non-Hispanic	2.43	2.96	2.99	4.00
Owner/other; Hispanic	2.00	3.06	2.81	5.66
Owner/other; black, non-Hispanic	2.53	3.46	2.90	6.11
Owner/other; white and other, non-Hispanic	2.26	3.45	2.03	3.10



service interruption prior to raking. The standard weights are not further adjusted while the alternative weights have different adjustments depending on the scheme.

#### 4. FINDINGS

As noted above, adjustment of the weights to reduce the bias increases the variability of the weights, thus increasing the variance of the estimates. Kish (1992) gives an approximate expression for this increase in variance arising from unequal weights. We call this expression for the increase in variance due to differential weights the variance inflation factor (*VIF*). The *VIF* can be written as

$$VIF = 1 + CV^2(\text{weights}) \quad (2)$$

where *CV* is the coefficient of variation of the weights.

Table 3 shows the *VIF* for the standard NHES:93 weights for each component. The SS&D component is broken down by the grade of the student, because youth were selected at different rates for these grade levels. The *VIF* for each of the components is about 1.4, indicating the variance is inflated by about 40 percent due to the variability in the standard weights. The *VIF* for the combined SS&D file is somewhat larger (1.5) because it includes youth who were sampled at different rates.

The other factors given in table 3 are the ratios of the *VIF* for the four alternative weights to the *VIF* for the standard weight. These ratios show how much greater the variances of estimates produced using the alternative weights are expected to be as compared to the variances of the standard NHES:93 weights.

Overall, the increase in variance due to the telephone interruption coverage adjustment are from 9 to 13 percent for schemes A1 and B1 in the SS&D component but up to 20 percent for the SR component. The ratios are larger for the schemes A2 and B2, ranging from 24 to 35 percent, with the largest ratio for Scheme A2 for the SR component. The larger ratios (hence *VIF*s) for the schemes based on interruptions of one month or more are a consequence of the larger and more variable factors shown in the second column of table 2. The ratios for the SR population are higher than the SS&D ratios.

#### 4.1 Coverage Bias Reduction

If estimates of the same characteristics as those produced from the NHES:93 were available from an independent source and these benchmark estimates were free of telephone coverage bias, then it would be possible to compare the five estimates to the benchmark. However, benchmarks comparable to the estimates from the two components of the NHES:93 do not exist and other methods are needed to assess the bias-reducing potential of the coverage adjustments.

Due to the lack of a benchmark, some model assumptions are required to assess the effectiveness of the adjustments. For this evaluation we assume that the adjustment procedures reduce the coverage bias. As a result of this assumption, the difference between the standard estimate and the adjusted estimate is considered an unbiased estimate of the decrease in the coverage bias resulting from using the procedures. Clearly, the coverage bias is not completely eliminated by any of the adjustment procedures. Even if the model were correct, the bias reductions from the data would still be subject to sampling error. Despite the problems with this assumption, this type of assumption is necessary to obtain some idea of the effectiveness of the adjustment. If the adjustment eliminates the bias, the mean square errors of the adjusted estimates are equal to the variances of the estimates, with no contribution from coverage bias. Therefore, the model assumption is favorable to the adjusted estimates, positing the adjusted estimates to be unbiased. The impact of this assumption is discussed critically after evidence of the effectiveness of the method is presented.

The estimate from each scheme can be compared to the standard NHES:93 estimate, and the difference between the standard estimate and the adjusted estimate is an estimate of the reduction in the coverage bias. With four adjusted estimates, four different estimates of bias reduction are possible. The estimated reduction in bias is

$$b_a = \hat{p}_s - \hat{p}_a, \quad (3)$$

where  $b_a$  is the estimated bias reduction using adjustment scheme  $a$  ( $a = A1, A2, B1, \text{ or } B2$ ),  $\hat{p}_s$  is the estimate of the proportion using the standard estimate, and  $\hat{p}_a$  is the estimated proportion using adjustment scheme  $a$ .

**Table 3**  
Ratios of Variance Inflation Factor Due to Coverage Adjustment

Component	Sample size	<i>VIF</i> * standard weight	Ratio of scheme's <i>VIF</i> to standard weight's <i>VIF</i>			
			Scheme A1	Scheme A2	Scheme B1	Scheme B2
School Readiness	10,888	1.36	1.20	1.35	1.16	1.26
School Safety and Discipline						
3rd through 5th graders	2,563	1.37	1.12	1.25	1.13	1.26
6th through 12th graders	10,117	1.39	1.13	1.27	1.09	1.24
3rd through 12th graders	12,680	1.49	1.12	1.26	1.11	1.25

\* *VIF* is the standard inflation factor. It is the coefficient of variation of the weights squared plus one.

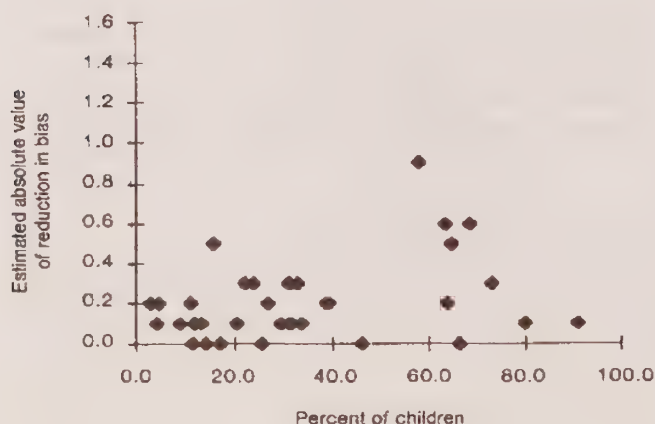
Source: U.S. Department of Education, National Center for Education Statistics, National Household Education Survey, spring 1993.

The estimated reductions in bias under each adjustment weighting scheme are given in table 4. Estimates for additional characteristics are given in Brick *et al.* (1996). The bias reductions in the standard estimate assume each adjustment scheme eliminates the coverage bias.

The bias reduction estimates for most of the items in Table 4 are less than one percent and consistent in direction across the schemes. Before summarizing the estimates, we must account for the fact that the total number of children is constant for all the estimates due to the raking of the estimates to the CPS totals. The fixed total number of children across response categories has two consequences: it creates a negative correlation in the estimated reduction in bias across response categories; and it gives a false impression of the number of independent pieces of information in the tabled values.

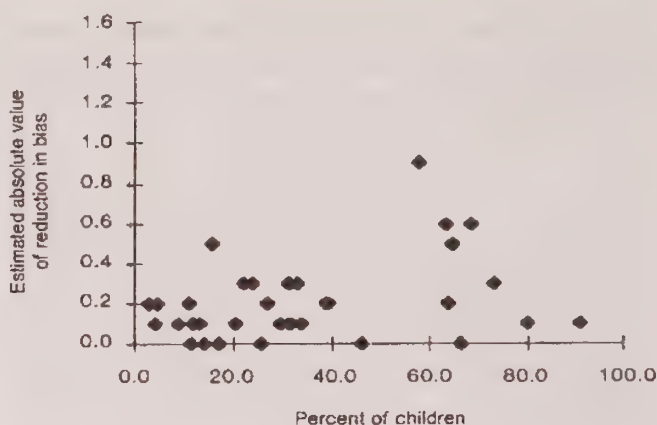
The approach taken to address to this problem in summarizing the bias estimates is to delete the estimate for one of the response categories for each item. The "no" response category for all items with "yes" and "no" response categories was deleted. For other types of variables, the response category with the smallest estimate was deleted.

Figure 2 presents the absolute value of the reduction in bias estimated using scheme A1 for the SR characteristics, and figure 3 is the same representation for the SS&D. These figures use all the estimates presented in Brick *et al.* (1996), rather than just those shown in table 4. For both components, the bias reductions are small. The largest absolute bias is 1.3 percent for SR and 0.9 percent for SS&D. The mean and median of the bias reductions and the absolute values of the bias reductions were also computed for each scheme and each component. For the SR component, the mean and median of the absolute value of the estimated bias reductions for the four schemes are between 0.2 and 0.4 percent. For the SS&D, the mean and median of the absolute values are between 0.1 and 0.3.



Source: U.S. Department of Education, National Center for Education Statistics, National Household Education Survey, spring 1993

Figure 2. Estimated reduction in absolute bias for School Readiness characteristics (scheme A1)



Source: U.S. Department of Education, National Center for Education Statistics, National Household Education Survey, spring 1993

Figure 3. Estimated reduction in absolute value of bias for School Safety and Discipline characteristics (scheme A1)

### Bias Ratio

The size of the absolute reduction in bias is not a very useful statistical measure of the impact of the bias because it does not take the magnitude of the sampling error of the estimate into account. Cochran (1977) discusses the impact on confidence intervals as the ratio of the bias to the sampling error varies. For each scheme the bias ratio is given by

$$r_a = \frac{b_a}{\text{se}(\hat{p}_s)}, \quad (4)$$

with the standard error of the standard estimate as the denominator. As the bias ratio increases, the chance of covering the population value departs significantly from the nominal confidence interval.

The bias ratios for selected characteristics are shown in Table 4. Many of the bias ratios for the SR items are large, even though the average and median ratios are near zero. Nearly half of the ratios for all the items examined are larger than 0.4 in absolute value. A ratio of 0.4 is large enough to reduce a nominal confidence interval from 95 percent to about 93 percent. For the SS&D items, the bias ratios are smaller, with only 15 percent of all the items having bias ratios greater than 0.4.

### 4.2 Mean Square Error

Since the variance is not an adequate measure of error for biased estimates, the mean square error of the estimates is used instead. The mean square error (MSE) is the sum of the variance and the square of the bias of the estimate.

The MSE can be estimated for the NHES:93 estimates by using the standard variance estimates and the bias reduction estimates presented above. The estimated MSE can be approximated as

$$\text{MSE}_a = \text{var}(\hat{p}_s) + b_a^2 \quad (5)$$



**Table 4**  
 Estimated Reduction in Bias and Bias Ratio for Selected Characteristics of the NHES:93

Characteristic	Standard estimate		Estimated reduction in bias				Bias ratio			
	Estimate	Standard error	Scheme A1	Scheme A2	Scheme B1	Scheme B2	Scheme A1	Scheme A2	Scheme B1	Scheme B2
<b>School Readiness (SR) population</b>										
Parental educational level										
Less than high school graduate	8.6	0.3	-1.7	-1.9	0.1	0.1	-5.7	-6.3	0.3	0.3
High school graduate or equivalent	33.9	0.8	0.4	0.3	-0.7	-1.0	0.5	0.4	-0.9	-1.3
Some college	57.5	0.7	1.3	1.6	0.6	0.9	1.9	2.3	0.9	1.3
Mother's employment status										
No mother in household	2.4	0.2	-0.1	-0.1	-0.1	-0.1	-0.5	-0.5	-0.5	-0.5
Employed 35 hours/week or more	34.3	0.5	0.5	0.8	0.2	0.5	1.0	1.6	0.4	1.0
Employed less than 35 hours/week	20.9	0.5	-0.1	-0.2	0.0	-0.2	-0.2	-0.4	0.0	-0.4
Seeking employment	6.6	0.4	0.0	-0.1	-0.1	-0.1	0.0	-0.3	-0.3	-0.3
Not in labor force	35.8	0.6	-0.4	-0.3	0.0	0.0	-0.7	-0.5	0.0	0.0
Father's employment status										
No father in household	26.3	0.5	-0.4	-0.6	0.0	-0.1	-0.8	-1.2	0.0	-0.2
Employed 35 hours/week or more	63.4	0.6	0.3	0.5	0.1	0.2	0.5	0.8	0.2	0.3
Employed less than 35 hours/week	3.8	0.3	0.0	-0.1	0.0	0.1	0.0	-0.3	0.0	0.3
Seeking employment	3.2	0.3	0.0	0.0	-0.1	-0.2	0.0	0.0	-0.3	-0.7
Not in labor force	3.3	0.2	0.1	0.2	0.0	0.1	0.5	1.0	0.0	0.5
Time since doctor visit for routine care										
Less than 1 year	84.1	0.4	0.4	0.4	0.2	0.1	1.0	1.0	0.5	0.2
Over 1 year	15.9	0.4	-0.4	-0.5	-0.2	-0.1	-1.0	-1.3	-0.5	-0.2
Birth weight										
5.5 pounds or less	93.3	0.3	-0.1	0.0	0.0	0.1	-0.3	0.0	0.0	0.3
Greater than 5.5 pounds	6.7	0.3	0.1	0.0	0.0	-0.1	0.3	0.0	0.0	-0.3
Child attending center-based program <sup>1</sup>										
Yes	52.6	0.8	0.9	0.3	0.8	0.6	1.1	0.4	1.0	0.8
No	47.4	0.8	-0.9	-0.3	-0.8	-0.6	-1.1	-0.4	-1.0	-0.8
Child ever attended center-based program <sup>1</sup>										
Yes	62.9	0.8	0.5	0.3	0.4	0.3	0.6	0.4	0.5	0.4
No	37.1	0.8	-0.5	-0.3	-0.4	-0.3	-0.6	-0.4	-0.5	-0.4
Attended center-based program prior to school <sup>2</sup>										
Yes	73.5	0.5	0.6	0.7	0.5	0.6	1.2	1.4	1.0	1.2
No	26.5	0.5	-0.6	-0.7	-0.5	-0.6	-1.2	-1.4	-1.0	-1.2
Women, Infant, and Children program participant <sup>1</sup>										
Yes	33.8	1.0	-0.6	-0.1	-0.8	-0.7	-0.6	-0.1	-0.8	-0.7
No	66.2	1.0	0.6	0.1	0.8	0.7	0.6	0.1	0.8	0.7
Free meal at school or center <sup>2</sup>										
Yes	35.8	0.6	-0.9	-1.1	-0.5	-0.5	-1.5	-1.8	-0.8	-0.8
No	64.2	0.6	0.9	1.1	0.5	0.5	1.5	1.8	0.8	0.8
Repeated kindergarten <sup>3</sup>										
Yes	5.7	0.4	-0.3	-0.5	-0.2	-0.2	-0.8	-1.3	-0.5	-0.5
No	94.3	0.4	0.3	0.5	0.2	0.2	0.7	1.3	0.5	0.5
<b>School Safety and Discipline (SS&amp;D) population</b>										
Parental educational level										
Less than high school graduate	9.4	0.5	-1.2	-1.3	-0.3	-0.6	-2.4	-2.6	-0.6	-1.2
High school graduate or equivalent	32.7	0.6	0.3	0.0	-0.2	-0.6	0.5	0.0	-0.3	-1.0
Some college	57.9	0.5	0.9	1.3	0.5	1.1	1.8	2.6	1.0	2.2
Mother's employment status										
No mother in household	3.5	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Employed 35 hours/week or more	46.2	0.5	0.0	0.1	-0.1	0.1	0.0	0.2	-0.2	0.2
Employed less than 35 hours/week	20.3	0.5	0.1	0.0	0.0	-0.1	0.2	0.0	0.0	-0.2
Seeking employment	4.5	0.3	-0.2	-0.2	-0.2	-0.2	-0.7	-0.7	-0.7	-0.7
Not in labor force	25.5	0.5	0.0	0.1	0.2	0.2	0.0	0.2	0.4	0.4

**Table 4**  
Estimated Reduction in Bias and Bias Ratio for Selected Characteristics of the NHES:93 – Concluded

Characteristic	Standard estimate		Estimated reduction in bias				Bias ratio			
	Estimate	Standard error	Scheme A1	Scheme A2	Scheme B1	Scheme B2	Scheme A1	Scheme A2	Scheme B1	Scheme B2
Father's employment status										
No father in household	26.8	0.6	-0.2	-0.2	-0.1	-0.2	-0.3	-0.3	-0.2	-0.3
Employed 35 hours/week or more	63.2	0.5	0.6	0.9	0.6	0.8	1.2	1.8	1.2	1.6
Employed less than 35 hours/week	3.1	0.2	-0.2	-0.2	-0.2	-0.2	-1.0	-1.0	-1.0	-1.0
Seeking employment	2.6	0.2	-0.2	-0.3	-0.2	-0.3	-1.0	-1.5	-1.0	-1.5
Not in labor force	4.3	0.3	-0.1	-0.1	-0.1	-0.1	-0.3	-0.3	-0.3	-0.3
School control										
Public	91.2	0.3	-0.1	-0.1	-0.1	-0.1	-0.3	-0.3	-0.3	-0.3
Private	8.8	0.3	0.1	0.1	0.1	0.1	0.3	0.3	0.3	0.3
Visitors required to sign in at school										
Yes	79.9	0.5	0.1	0.4	0.0	0.2	0.2	0.8	0.0	0.4
No	20.1	0.5	-0.1	-0.4	0.0	-0.2	-0.2	-0.8	0.0	-0.4
Had drug or alcohol ed program this year										
Yes	68.5	0.7	0.6	0.8	0.7	0.9	0.9	1.1	1.0	1.3
No	31.5	0.7	-0.6	-0.8	-0.7	-0.9	-0.9	-1.1	-1.0	-1.3
Students in fighting gangs at school <sup>4</sup>										
Yes	22.3	0.5	-0.3	-0.4	-0.3	-0.5	-0.6	-0.8	-0.6	-1.0
No	77.7	0.5	0.3	0.4	0.3	0.5	0.6	0.8	0.6	1.0
Ease of obtaining marijuana at school <sup>4</sup>										
Very or fairly easy	39.2	0.6	-0.2	-0.3	-0.2	-0.3	-0.3	-0.5	-0.3	-0.5
Hard	29.7	0.5	0.1	0.1	0.2	0.2	0.2	0.2	0.4	0.4
Nearly impossible	31.1	0.6	0.1	0.1	0.0	0.1	0.2	0.2	0.0	0.2
Fear of incident of crime at school										
None	66.1	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Fear of theft or robbery <sup>5</sup>	11.9	0.5	-0.1	-0.2	0.0	-0.2	-0.2	-0.4	0.0	-0.4
Fear of bullying or assault <sup>5</sup>	8.6	0.3	-0.1	-0.1	-0.1	-0.1	-0.3	-0.3	-0.3	-0.3
Fear of two or more types of incidents <sup>5</sup>	13.3	0.5	0.1	0.3	0.1	0.2	0.2	0.6	0.2	0.4
Knowledge of crime at school										
None	38.7	0.6	0.2	0.1	0.2	0.1	0.3	0.2	0.3	0.2
Fear of theft or robbery <sup>5</sup>	14.1	0.5	0.2	0.3	0.2	0.3	0.4	0.6	0.4	0.6
Fear of bullying or assault <sup>5</sup>	15.6	0.4	-0.5	-0.4	-0.4	-0.4	-1.3	-1.0	-1.0	-1.0
Fear of two or more types of incidents <sup>5</sup>	31.6	0.6	0.1	0.0	0.0	0.0	0.2	0.0	0.0	0.0
Victimization by crime										
Not victimized	73.0	0.5	0.3	0.2	0.3	0.2	0.6	0.4	0.6	0.4
Victim of theft or robbery <sup>5</sup>	10.9	0.3	-0.2	-0.1	-0.1	0.0	-0.7	-0.3	-0.3	0.0
Victim of bullying or assault <sup>5</sup>	8.9	0.3	-0.1	0.0	-0.2	-0.1	-0.3	0.0	-0.7	-0.3
Victim of two or more types of incidents <sup>5</sup>	7.2	0.3	0.0	0.0	0.0	-0.1	0.0	0.0	0.0	-0.3
Witnessed crime at school										
None	63.8	0.8	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
Witnessed robbery <sup>6</sup>	0.6	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Witnessed bullying or assault <sup>6</sup>	24.1	0.8	-0.3	-0.3	-0.3	-0.3	-0.4	-0.4	-0.4	-0.4
Witnessed two or more types of incidents	11.4	0.4	0.0	0.1	0.0	0.0	0.0	0.2	0.0	0.0

<sup>1</sup> Applies to preschoolers only.

<sup>2</sup> Applies to all children except preschoolers.

<sup>3</sup> Applies to children in primary school only.

<sup>4</sup> Applies to students in grades 6 through 12 only.

<sup>5</sup> For the fear of incident, knowledge of crime, and victimized by crime variables, the second response category is used if either theft or robbery was reported but not both, the third response category is used if either bullying or assault was reported but not both.

<sup>6</sup> This response category is used if either bullying or assault was reported, but not both, was reported.

Note: Percents may not add to 100 because of rounding.

Source: U.S. Department of Education, National Center for Education Statistics, National Household Education Survey, spring 1993.



where  $\hat{p}_s$  is the estimated proportion under the standard approach and  $b_a$  is the reduction in bias under scheme  $a$ . Because of the high correlation in the estimates of the bias from the four adjustment schemes, only the mean square errors for scheme A1 were computed. In Brick *et al.* (1996), the estimates using other schemes are shown to have negligible effects.

The mean square errors of the adjusted estimates are now contrasted with the variability in the standard NHES:93 estimates. The variance increase from adjusting the weights using the telephone service interruption data was expressed as a *VIF* in table 3. Multiplying the variance estimates of the standard estimates by the appropriate adjustment factor yields an approximate variance for the adjusted (presumably unbiased) estimates which are then compared to the mean square error of the standard estimates.

To aid in comparing the weighting procedures, ratios of the variance of the adjusted estimate to the mean square error for the standard estimate were tabulated (see Brick *et al.* 1996). The ratio is called the mean square ratio and can be written as

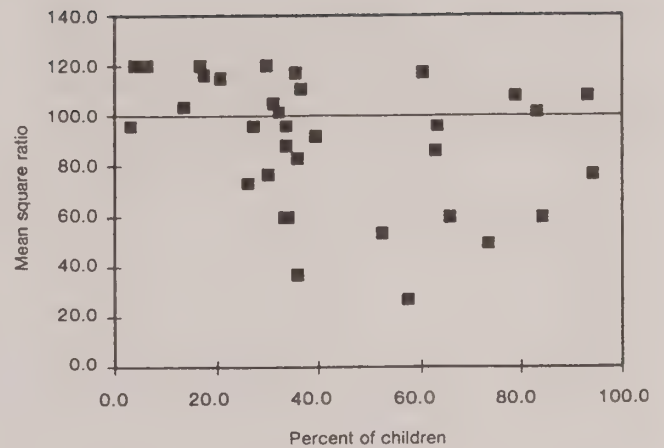
$$\text{msr}_a(\hat{p}) = \frac{100 \times \text{relativeVIF}_a \times \text{var}(\hat{p}_s)}{\text{mse}_{A1}(\hat{p})} \quad (6)$$

Note that the mean square error is derived using the bias estimated from scheme A1 only, but it is used to compute the mean square ratios for all four schemes. As noted above, this simplification does not have much effect on the mean square ratios because the bias estimates are approximately the same across schemes.

The mean square ratios include contributions from the bias (in the mean square error estimates) and the variance (in the *VIF*). When the mean square ratio is 100, the variance of the adjusted estimate is exactly equal to the mean square error of the biased, standard estimate. A ratio less than 100 indicates that the bias reduction of the adjustment is greater than the variance increase that comes with it. A mean square ratio over 100 means that the variance increase associated with the adjustment is greater than the bias reduction.

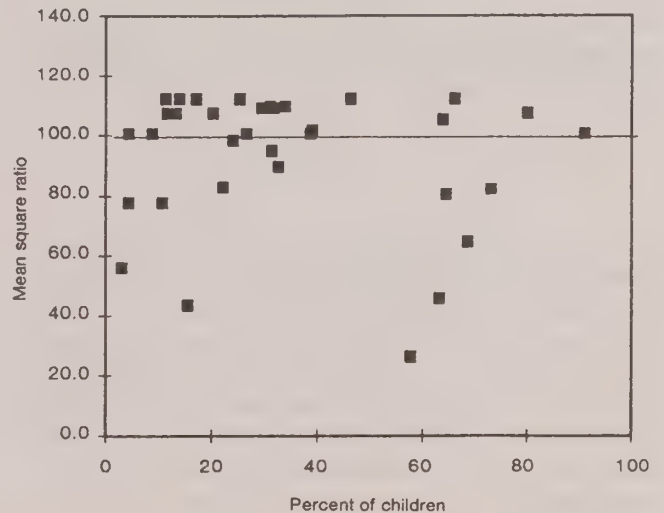
Figures 4 and 5 graphically present the msr for the two component surveys using scheme A1. In addition, Table 5 shows summary statistics for the msr for all four adjustment schemes. The distributions of mean square ratios for both components are very similar with the mean square ratios slightly lower for the SR component. The medians for schemes A1 and B1 (those based on interruptions of one week or more) are near the break-even point of 100. The means for these schemes are close to 90 and the figures confirm that the difference between the mean and medians is due to the skewed distributions of the mean square ratios.

A striking feature of the distributions of the mean square ratios for schemes A1 and B1 is the size of the ratios at the extremes of the distribution. The maximum mean square ratios for both components is 120, while some ratios are as small as 26. This means the maximum increase in the mean square error of the estimates is 20 percent, while the reductions in



Source: U.S. Department of Education, National Center for Education Statistics, National Household Education Survey, spring 1993

Figure 4. Estimated mean square ratios for selected School Readiness items (scheme A1)



Source: U.S. Department of Education, National Center for Education Statistics, National Household Education Survey, spring 1993

Figure 5. Estimated mean square ratios for selected School Safety and Discipline items (scheme A1)

mean square error for a number of other estimates are quite large. Thus, the penalty associated with adjusting even when the estimate is not biased is modest, but the benefits of adjusting when it is needed are impressive.

The distributions for the mean square ratios for schemes A1 and B1 are very similar, and the choice of which of these schemes should be used may be determined by nonstatistical issues, such as availability of data and the other types of adjustments required in the survey. The mean square ratios show that the adjusted weights reduce the mean square error for about half the estimates considered below those derived from the standard weights. The distributions of the mean square ratios for schemes A2 and B2 (those based on interruptions of 1 month or more) have medians and means

**Table 5**

Summaries of Distribution of Mean Square Ratios for Selected Characteristics of School Readiness and School Safety and Discipline Components

	Adjustment scheme			
	A1	A2	B1	B2
<b>School Readiness</b>				
Mean	89.8	101.0	86.8	94.2
Median	96.0	108.0	92.8	100.8
Minimum	27.0	30.3	26.1	28.3
Maximum	120.0	135.0	116.0	126.0
<b>School Safety and Discipline</b>				
Mean	93.3	104.9	92.2	103.9
Median	100.8	113.4	99.9	112.5
Minimum	26.4	29.7	26.2	29.5
Maximum	112.0	126.0	111.0	125.0

**Source:** U.S. Department of Education, National Center for Education Statistics, National Household Education Survey, spring 1993.

that are greater than 100. Essentially, these mean square ratios are shifted upward when compared with those of schemes A1 and B1, and are not recommended.

## 5. CONCLUSIONS

If the percentage of the target population living in non-telephone households is relatively large and the characteristics of those persons are different from those who live in telephone households, then the estimates may be susceptible to significant coverage bias. One method of addressing this problem without resorting to other modes of data collection is to adjust the weights to reduce the coverage bias. In this study, the weights for persons in households reporting an interruption in telephone service were increased to account for those without telephones.

The bias reduction estimates computed under the assumed model showed that the coverage adjustments for the SR component improved some of the estimates substantially, and did not do much harm to any statistics. The bias reduction estimates from the SS&D component, on the other hand, were not as substantively important. The adjustments reduced bias for both components, but they also increased the variability of the estimates. The distributions of the mean square ratios show that about half the estimates could be improved using the telephone service interruption adjustments. Furthermore, even for those estimates that were less accurate due to the variance increases associated with the differential weights, the magnitude of the increases were not large. In other words, the penalty for adjusting when it did not reduce the coverage bias was not very great. These findings suggest that the adjustments should be seriously considered.

The alternative weighting schemes performed differently with respect to the mean square ratios. The schemes based on

interruptions of telephone service of one week or more were better than the schemes based on interruptions of one month or more. The bias adjustments resulting from using educational attainment by race/ethnicity categories were roughly equivalent to those using tenure by race/ethnicity.

The size of the sample is a relevant factor that should be taken into account when evaluating the use of the telephone service interruption adjustment. Bias ratios increase with the sample size because the bias is not affected while the sampling error of the estimate (the denominator of the bias ratio) decreases. Thus, the adjustments should be more beneficial in surveys with large sample sizes where the bias ratios might be expected to be large.

While the results of this study suggest that the adjustments could be useful for many estimates from telephone surveys, confirmation is needed before the adjustments are recommended. As discussed earlier, the estimates of the mean square errors in this study were based on the assumption that the adjusted estimates eliminated the bias of the estimates. This model assumption could not be verified because of the lack of benchmark data for comparison. The assumed model is very beneficial to the adjusted estimates in the sense that it results in lower bounds on the mean square errors for the adjusted estimates. Thus, the findings of this study should be taken as an indication that adjustment using data on interruptions in telephone service is a feasible method, but requires further study and evaluation.

The questions about interruptions in telephone service were recently added to the National Health Interview Survey, a survey conducted by the Census Bureau for the National Center for Health Statistics. The findings from this survey should be very useful in evaluating this method because the survey covers households without telephones by in-person interviews, eliminating the need for the critical model assumption used in this study.

## ACKNOWLEDGEMENTS

The authors would like to thank the referees and editor for comments that substantially improved the methods and presentation of the material.

## REFERENCES

- BRICK, J., BURKE, J., and WEST, J. (1992). Telephone Undercoverage Bias of 14- to 21-year-olds and 3- to 5-year-olds. National Household Education Survey Technical Report No. 2, Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, NCES 92-101.
- BRICK, J., KEETER, S., WAKSBERG, J., and BELL, B. (1996). Adjusting for Coverage Bias Using Telephone Service Interruption Data. National Household Education Survey Technical Report, Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, NCES 96-336.



- COCHRAN, W. (1977). *Sampling Techniques*. New York: John Wiley and Sons, 12-15.
- CONVERSE, J., and PRESSER, S. (1986). *Survey Questions, Handcrafting the Standardized Questionnaire*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-063. Beverly Hills: Sage Publishers.
- FEDERAL COMMUNICATIONS COMMISSION (1988). Monitoring Report: CC Docket No. 87-339. Prepared by the Staff of the Federal-State Joint Board in CC Docket No. 80-286, Washington DC.
- KALTON, G., and KASPRZYK, D. (1986). The treatment of missing survey data. *Survey Methodology*, 12, 1-16.
- KEETER, S. (1995). Estimating noncoverage bias from a phone survey. *Public Opinion Quarterly*, 59, 196-217.
- KISH, L. (1992). Weighting for unequal  $P_i$ . *Journal of Official Statistics*, 8, 183-200.
- MASSEY, J., and BOTMAN, S. (1988). Weighting adjustments for random digit dialed surveys, Chapter 9 in *Telephone Survey Methodology*. Eds. Groves, Biemer, Lyberg, Massey, Nicholls, and Waksberg. New York: John Wiley and Sons, 143-160.
- POLITZ, A., and SIMMONS, W. (1949). An attempt to get the 'not at homes' into the sample without callbacks. *Journal of the American Statistical Association*, 44, 9-31.
- THORNBERRY, O., and MASSEY, J. (1988). Trends in United States telephone coverage across time and subgroups, Chapter 3 in *Telephone Survey Methodology*. Eds. Groves, Biemer, Lyberg, Massey, Nicholls, and Waksberg. New York: John Wiley and Sons, 25-50.





# Optimal Sample Redesign Under GREG in Skewed Populations With Application

GURUPDESH S. PANDHER <sup>1</sup>

## ABSTRACT

Within a survey re-engineering context, the combined methodology developed in the paper addresses the problem of finding the minimal sample size for the generalized regression estimator in skewed survey populations (*e.g.*, business, institutional, agriculture populations). Three components necessary in identifying an efficient sample redesign strategy involve i) constructing an efficient partitioning between the "take-all" and "sampled" groups, ii) identifying an efficient sample selection scheme, and iii) finding the minimal sample size required to meet the desired precision constraint(s). A scheme named the "Transfer Algorithm" is devised to address the first issue (Pandher 1995) and is integrated with the other two components to arrive at a combined iterative procedure that converges to a globally minimal sample size and population partitioning under the imposed precision constraint. An equivalence result is obtained allowing the solution to the proposed algorithm to be alternatively determined in terms of simple quantities computable directly from the population auxiliary data. Results from the application of the proposed sample redesign methodology to the Local Government Survey in Ontario are reported. A 52% reduction in the total sample size is achieved for the regression estimator of the total at a minimum coefficient of variation of 2%.

**KEY WORDS:** Minimal sample size; Optimal sample selection; Precision constraint; Sampled group; Take-all group.

## 1. INTRODUCTION

In many survey situations additional information is available on all population units before the survey is undertaken. This auxiliary information is frequently useful in devising a more efficient sample design and estimation strategy. In a survey redesign context, the most optimal strategy holds the promise of offering the largest reduction in survey costs by requiring the lowest sample size necessary to meet the desired precision constraint on the estimates. In repeat surveys of skewed populations, an efficient sample design and estimation strategy may be realized by exploiting a) the correlation structure between the size-based auxiliary information  $x$  (*e.g.*, population of municipality, employees in a firm, farm acreage) and the survey variables  $y$  (*e.g.*, municipality expenditures, value of shipments, farm yield) and b) the variance relationship between the survey variable and the auxiliary size information.

In this paper, a comprehensive sample redesign methodology is developed for skewed populations with the ultimate objective of bringing about maximal reductions in the current sample size while ensuring a desired level of precision for the generalized regression estimator of the total. This work was motivated by the redesign of the Local Government Finance Survey (LGFS) conducted by Statistics Canada's Public Institutions Division. Financial information (*e.g.*, revenues, expenditures, debt, *etc.*) obtained from local government units is used in the estimation and publication of financial statistics on a provincial and national basis.

Although the work presented in this paper is motivated by a concrete application, the sample design methodology devised applies generally to all surveys based on skewed populations (*e.g.*, agricultural, business, and institutional surveys).

In identifying an efficient new sample design, the overall methodology addresses and integrates the solution to three problems:

### 1) Creation of the "Take-all" and "Sampled Groups"

Since the variability of the survey response  $y_k$  tends to increase with the size of the unit  $x_k$ , it is common in skewed populations to sample the largest  $x$ -valued units with certainty in order to improve the efficiency of the population estimators. The demarcation of the population into the non-overlapping "take-all"  $U_a = \{1, \dots, N_a\}$  and "sampled" groups  $U_b = \{1, \dots, N_b\}$  is obtained through a new scheme named the "Transfer Algorithm".

### 2) Choosing an Efficient Sample Selection Scheme

Let  $p(s; \lambda) = (p_a(s_a), p_b(s_b; \lambda))$  represent the complete sample design where the sample design parameter  $\lambda$  determines the type of sample selection implemented in the sampled group  $U_b$ . The sample inclusion probabilities due to  $p_b(s_b; \lambda)$  may be expressed as  $\pi_k(\lambda) = n_b (x_k^{\lambda/2} / \sum_{j \in U_b} x_j^{\lambda/2})$ ,  $k \in U_b$ . Note that the parameter  $\lambda$  defines a broad class of sample designs with SRS ( $\lambda = 0$ ) and pps ( $\lambda = 2$ ) as particular cases. Design optimality results (Godambe and Joshi 1965) allow the identification of the most optimal value for the sample design parameter  $\lambda$ .

<sup>1</sup> Gurupdes S. Pandher, Survey Analysis and Methods Development Section, Household Survey Methods Division, Methodology Branch, Statistics Canada, 16th Floor, R.H. Coats Building, Ottawa, Ontario, Canada, K1A 0T6.

### 3) Minimal Sample Size Determination

The third component of the overall methodology is aimed at finding the minimal sample size required to meet the imposed precision constraints for the estimator.

The combined procedure devised integrates these components to allow a new globally minimal sample size and optimal population partitioning to be determined under a flexible range of sample selection strategies (*e.g.*, SRS, pps, generalized pps). Firstly, the “Transfer Algorithm” is proposed which finds an optimal population allocation between the take-all and sampled population groups in the sense of minimizing the variance of the generalized regression estimator (GREG) of the total. Desirable mathematical properties of this algorithm such as existence and optimality of solution along with an equivalence result were established in Pandher (1995). The equivalence result allows the solution to be determined in terms of simple quantities computable directly from the population auxiliary data.

The Transfer Algorithm in then synthesized iteratively with the sample size determination step to find the minimal sample size needed to satisfy the imposed precision constraints through an iterative procedure. The combined methodology produces a sequence of sample sizes and population partitionings which converge to a globally optimal solution where further reductions in the sample size are not possible given the imposed precision constraint. An application of the procedure is given for Ontario using provincial data from the Local Government Finance Survey.

Lavallée and Hidirolou (1988), Hidirolou and Srinath (1993) (subsequently denoted as L&H and H&S, respectively), and Glasser (1962) have proposed alternative methodologies for constructing the take-all and sampled groups within the context of stratified SRS design. The proposed approach differs from other methods in three respects. Firstly, the population demarcation is obtained under a flexible range of sample selection strategies (*e.g.*, SRS, pps, generalized pps). Secondly, the criterion for constructing the population demarcation is based on minimizing the variance of the GREG estimator of the total under the desired sample selection strategy (Glasser and L&H base their allocation on minimizing the within-stratum sum-of-squares  $x$ ; H&S use the total regression sum-of-squares under a regression model with a compulsory intercept assuming SRS). Thirdly, the proposed methodology explicitly captures the size-induced heteroscedasticity present in skewed survey populations which has been ignored in other frameworks.

Lastly, it is useful to qualify the sense in which the term “optimal” is used. Since, the redesign uses auxiliary information from a previous cycle of the survey to estimate the design parameters, there is a level of sub-optimality introduced in the redesign methodology by this lag. But as a practical matter, using the data from the most recent survey is the best that can be done. Once the design parameters have been estimated or are known however, the cut-offs and sample sizes required to achieve the desired precision yield the lowest anticipated design variance given that the estimates

are true (or close to it). It is therefore, in this sense that the word “optimal” is used.

## 2. SURVEY FRAMEWORK

The model assisted survey framework is adopted for the skewed population whose auxiliary and survey characteristics are denoted by  $C_U = \{(x_1, y_1), \dots, (x_N, y_N)\}$ . In this framework, underlying the class of generalized regression estimators for the population total are regression models (Särndal 1992, p. 255) exploiting the correlation between the survey variables  $y$  and the auxiliary covariates  $x$ . Different model assumptions on the deterministic and stochastic components of the underlying model lead to different regression estimators for the population total. For example, a ratio-form heteroscedastic model

$$y_k = \beta x_k + \epsilon_k, \quad (2.1)$$

with the error  $\epsilon_k \sim (0, \sigma_k^2)$  and the variance structure given by  $\sigma_k^2 = c x_k^\gamma$  ( $\gamma$  is the heteroscedasticity parameter) leads to the following GREG estimator:

$$\hat{t}_{Rb} = \sum_{U_b} x_k \hat{B} + \sum_{s_b} \frac{(y_k - x_k \hat{B})}{\pi_k} \quad (2.2)$$

where  $\hat{B} = (\sum_s y_k / \pi_k) / (\sum_s x_k / \pi_k)$  is the sample-based probability weighted estimate of the population regression parameter  $B$ .

Given this estimation framework, the total across both groups  $t = t_a + t_b$  is estimated by  $\hat{t} = \hat{t}_a + \hat{t}_{Rb}$  where  $\hat{t}_a = \sum_{U_a} y_k$  since all units are sampled in the take-all group and  $\hat{t}_{Rb}$  is the GREG estimator under the relevant model. The anticipated variance of  $\hat{t}_{Rb}$  (defined as the variance with respect to both the design and the model, denoted  $p$  and  $\xi$ , respectively) is expressible as

$$V(\hat{t}_{Rb}) \equiv E_\xi V_p(\hat{t}_{Rb}) = \sum_{k \in U_b} \left( \frac{1}{\pi_k} - 1 \right) \sigma_k^2. \quad (2.3)$$

Furthermore, if  $\sigma_k^2$  depends on the auxiliary measure  $x_k$  according to the formulation  $\sigma_k^2 = c x_k^\gamma$  (2.4), then design optimality (Godambe and Joshi 1965) implies that the optimal sample inclusion probabilities are  $\pi_k^*(\gamma) \propto x_k^{\gamma/2}$ ,  $k \in U_b$ . Therefore, the sample design  $p_b^*(s_b; \lambda = \gamma)$  in the sampled sub-population, defining the first order inclusion probabilities  $\pi_k^*(\gamma) = n(x_k^{\gamma/2} / \sum_{U_b} x_j^{\gamma/2})$ ,  $k \in U_b$ , minimizes the anticipated variance  $V(\hat{t}_{Rb})$ .

In the model assisted framework used in this paper, the auxiliary measure  $x_k$  is assumed to be a scalar. As noted by a referee, the more general case where  $x_k$  is a vector could be handled by fitting the appropriate parametric relationship  $\sigma_k^2 = f(x_{k1}, \dots, x_{kq})$  and using the estimated  $\hat{\sigma}_k$  in lieu of  $\sigma_k$  in defining the inclusion probabilities. The approach for the multivariate  $x_k$  seems intuitively sound and is mentioned here for completeness but requires further study and investigation.



Three methods for estimating the heteroscedasticity parameter  $\gamma$  from past survey data called the "Least Squares Method", the "Maximum Likelihood Method", and the "Graphical Method" are described in Appendix A of Pandher (1995).

### 3. TRANSFER ALGORITHM

In this section, an iterative scheme named the "Transfer Algorithm" is proposed to determine the optimal demarcation between the take-all and sampled sub-populations under the sample design  $p(s; \lambda)$ . The criterion for this construction is based on finding a population partitioning minimizing the estimated anticipated variance of  $\hat{t}_{Rb}$ . An equivalence result from Pandher (1995) is used to find an alternative and simpler method of solution based entirely on quantities defined on the auxiliary population data.

The proposed scheme for constructing the take-all and sampled sub-populations,  $U_a$  and  $U_b$ , respectively, is based on the following idea. Initially, place all population units in the sampled group, labelling it  $U_b^{(0)}$  (the superscript  $l$  represents the iteration cycle). Hence, the take-all group is an empty set  $U_a^{(0)} = \{\emptyset\}$ . The resulting population and sample size allocation at  $l = 0$  is given by  $N_a^{(0)} = 0$ ,  $n_a^{(0)} = 0$ ,  $N_b^{(0)} = N$ , and  $n_b^{(0)} = n_0$  where  $n_0$  is the current sample size.

In a repeat survey setting, the variances  $\sigma_k^2$  in (2.3) can be empirically modelled using the relation  $\sigma_k^2 = c x_k^\gamma$  (2.4) where  $\gamma$  and  $c$  are estimated from previous sample data as mentioned before. Using the estimated version of (2.4) in (2.3) yields the following estimator for  $V^{(l)}(\hat{t}_{Rb}; \cdot)$ :

$$\hat{V}^{(l)}(\hat{t}_R; \lambda, N_b^{(l)}, n_b^{(l)}) = \sum_{k \in U_b^{(l)}} \left( \frac{1}{\pi_k(\lambda)} - 1 \right) \hat{c} x_k^\gamma \quad (3.1)$$

where the largest  $l$   $x$ -valued units have been removed from  $U_b^{(0)}$ . Note that  $\lambda$  is used here to parameterize the sample design to allow greater generality when  $\lambda \neq \gamma$ .

In the iterative algorithm, we start initially with all population units placed in  $U_b^{(0)}$ . Then for each iteration  $l$ ,  $0 \leq l < n$ , the largest  $l+1$   $x$ -valued unit  $x_{(N-l-1)}$  is transferred from  $U_b^{(l)}$  to  $U_a^{(l)}$  and the difference

$$\Delta(l) = \hat{V}^{(l+1)}(\hat{t}_{Rb}; \lambda, N-l-1, n-l-1) - \hat{V}^{(l)}(\hat{t}_{Rb}; \lambda, N-l, n-l) \quad (3.2)$$

is computed. Negative values of  $\Delta(l)$  mean that the transfer of the unit corresponding to the ordered value  $x_{(N-l-1)}$  lead to a decrease in the variance. Moreover, such transfers continue to result in a reduction in the variance of  $\hat{t}_{Rb}$  as long as  $\Delta(l) < 0$ . In general, for any iteration  $l$ , the relationship between the population and sample size allocations is described by the following relations:  $N_b^{(l)} = N-l$ ,  $n_b^{(l)} = n-l$ , and  $N_a^{(l)} = n_a^{(l)} = l$ . These relations hold because the overall population and sample sizes must remain constant ( $N = N_a^{(l)} + N_b^{(l)}$  and  $n = n_a^{(l)} + n_b^{(l)}$ ) for all iterations.

The solution is also constrained by the condition  $\pi_k(\lambda) < 1$ ,  $k \in U_b^{(l^*)}$ . Let  $l^*(\lambda)$ ,  $0 \leq l^* < n$ , represent the solution to the Transfer Algorithm. Given the discussion above, the solution to the Transfer Algorithm under the sample design  $p(s; \lambda)$  may be formulated as

$$l^*(\lambda) = \min_l \{l: [\pi_{(N-l)}(\lambda) < 1] \quad \text{and}$$

$$\hat{\Delta}(l) = [\hat{V}^{(l+1)}(\hat{t}_{Rb}; \lambda) - \hat{V}^{(l)}(\hat{t}_{Rb}; \lambda)] \geq 0, 0 \leq l < n\}. \quad (3.3)$$

The optimal population allocation to the take-all group  $U_a^{(l^*)}$  is then given by the population units coinciding with the  $l^*$  ordered units transferred to the take-all auxiliary vector  $X_a^{(l^*)} = (x_{(N-l^*)}, x_{(N-l^*-1)}, \dots, x_{(N)})$ ; correspondingly the sampled group  $U_b^{(l^*)}$  consists of the units corresponding to  $X_b^{(l^*)} = (x_{(1)}, x_{(2)}, \dots, x_{(N-l^*-1)})$ .

Transferring a unit from  $U_b^{(l)}$  to  $U_a^{(l)}$  causes two opposite effects on the variance  $V^{(l)}(\hat{t}_{Rb}; \cdot)$ . The reduction in the population size ( $N_b^{(l+1)} = N_b^{(l)} - 1$ ) has the impact of decreasing the variance, while the equivalent reduction in the sample size ( $n_b^{(l+1)} = n_b^{(l)} - 1$ ) has the reverse effect of increasing  $V^{(l)}(\hat{t}_{Rb}; \cdot)$ . Somewhere in this process, a critical value  $l^*$ ,  $0 \leq l^* < n$ , exists which gives the optimal breakdown  $\{U_a^{(l^*)}, U_b^{(l^*)}\}$ . Moreover, in Theorem 3 of Pandher (1995), it is shown that as long as the conditions  $(x_{(N-l)}^{\lambda/2} - x_{(N-l-1)}^{\lambda/2}) \geq 0$  and  $(x_{(N-l)}^{\gamma-\lambda/2} - x_{(N-l-1)}^{\gamma-\lambda/2}) \geq 0$ ,  $0 \leq l < n$ , hold, a solution to the Transfer Algorithm exists and that the system remains stable (optimal) upon reaching  $l^*$ . Stability further implies that the solution is optimal since the conditions leading to the solution do not change in the range  $l^* \leq l < n$ . These two properties may be more precisely defined as follows:

Existence:  $\exists l^*$ ,  $0 \leq l^* < n$ , such that  $V^{(l^*+1)} - V^{(l^*)} \geq 0$  and  $\pi_{(N-l^*)}^{(l^*)} < 1$ .

Stability: If  $V^{(l^*+1)} - V^{(l^*)} \geq 0$ , then  $V^{(l+1)} - V^{(l)} \geq 0$  and  $\pi_{(N-l)}^{(l)} < 1$  for  $0 \leq l^* < l < n$ .

An example of the application of the Transfer Algorithm to the LGF survey population of local municipalities in Ontario (with  $N = 793$ ,  $n = 108$ ,  $\gamma = 2$ , and  $\lambda = 1$ ) is given in Figure 1. The curves are plotted for  $l > 8$  because in the interval  $0 < l \leq 8$ , the first condition of (3.3), namely  $[\pi_{(N-l)}(\lambda) < 1]$ , is not satisfied. The minimum value of  $\hat{V}^{(l)}(\hat{t}_{Rb})$  is achieved at  $l^* = 57$  where  $\Delta(l^*) = \hat{V}^{(l^*+1)} - \hat{V}^{(l^*)} \geq 0$ .

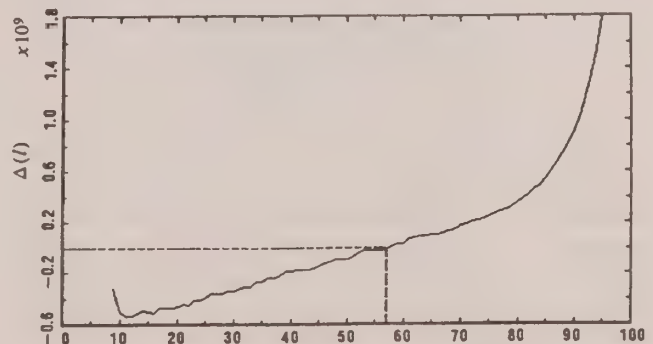


Figure 1. Changes in variance of regression estimator ( $\lambda = 1$ ):  $\Delta(l) = V^{(l+1)}(t; 1, N-l-1, n-l-1) - V^{(l)}(t; 1, N-l, n-l)$

Theorem 2 from the complete paper is an important result which allows the solution to the Transfer Algorithm to be equivalently expressed in terms of simpler quantities based on the auxiliary data. A brief sketch of the development of this theorem is given in the Appendix.

### Theorem 2. Equivalent Solution to the Transfer Algorithm

The solution  $l^*(\lambda)$  to the Transfer Algorithm stated in (3.3) in terms of  $V^{(l)} - V^{(l-1)}$  and  $\pi_{(N-l)}^{(l)}(\lambda)$  may also be equivalently expressed as

$$l^*(\lambda) = \begin{cases} \min_l \{l : n - l \leq R(l; \gamma - \lambda/2), 0 \leq l < n\}, 0 \leq \lambda < \gamma \\ \min_l \{l : n - l \leq R(l; \gamma/2), 0 \leq l < n\}, \lambda = \gamma \\ \min_l \{l : n - l \leq R(l; \lambda/2), 0 \leq l < n\}, \gamma < \lambda \leq 2\gamma \end{cases}$$

where  $R(l; \gamma - \lambda/2) = \sum_{k=1}^{N-l} x_{(k)}^{\gamma-\lambda/2} / x_{(N-l)}^{\gamma-\lambda/2}$  and  $R(l; \lambda/2) = \sum_{k=1}^{N-l} x_{(k)}^{\lambda/2} / x_{(N-l)}^{\lambda/2}$  define the critical values.

This use of this theorem to find the optimal population allocation is illustrated graphically in Figure 2 (Ontario data). In this case,  $0 \leq \lambda < \gamma$ , and the solution is determined by the behaviour of the functions  $R(l; \gamma - \lambda/2)$  (the lower curve in the graph) and  $n - l$ . The same solution  $l^* = 57$  is obtained as before.

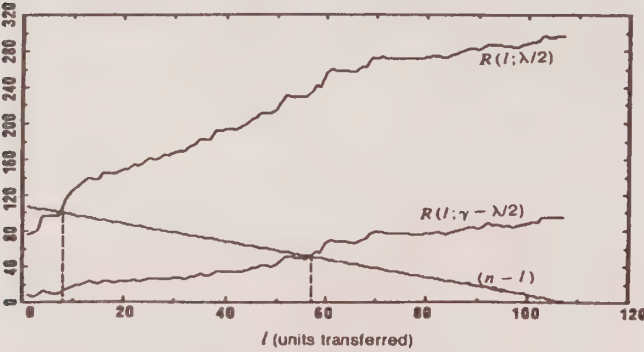


Figure 2. Use of  $R(l; \gamma - \lambda/2)$ ,  $R(l; \lambda/2)$ , and  $(n - l)$  to construct optimal take-all/sampled groups (Ontario)

## 4. SAMPLE SIZE DETERMINATION AND COMBINED ITERATIVE PROCEDURE

Given a sample design  $p(s, \lambda)$ ,  $0 \leq \lambda \leq 2\gamma$ , with sample size  $n$ , the Transfer Algorithm yields an optimal construction of the take-all and sampled sub-populations,  $U_a^*(l^*)$  and  $U_b^*(l^*)$ , respectively. Next, an expression for finding the minimal sample size is obtained which meets the imposed precision constraint – expressed in terms of the coefficient of variation  $CV_{\min}$ . The sample determination step is then integrated with the Transfer Algorithm to develop a combined procedure which allows the survey designer to find the globally minimal sample size and optimal population partitioning.

### 4.1 Expression for New Sample Size

Let  $q$  represent the iteration cycle for the combined procedure and  $n_q^* = n_{aq}^* + n_{bq}^*$  denote the total minimal sample size required to satisfy the precision constraint. Given the sample design  $p_q(s, \lambda, l_q^*(\lambda, n_q^*))$ , current sample size  $n_q$ , and the population partitioning  $\{U_{aq}^*(l_q^*), U_{bq}^*(l_q^*)\}$ , the precision constraint for  $\hat{t}_R = \hat{t}_a + \hat{t}_{Rb}$  may be stated formally as

$$CV_{\min} \geq \frac{\hat{V}_q^{1/2}(\hat{t}_{Rb}; \lambda, N - l_q^*, n_q - l_q^*)}{\hat{t}_R}. \quad (4.1)$$

Solving this inequality for  $n_{bq}^*$  gives the following expression for the minimal sample size needed in the sampled group  $U_{bq}^*(l_q^*)$  to meet the precision constraint:

$$n_{bq}^* = n_q^* - l_q^*(n_q) = \frac{X(l_q^*, \lambda/2) X(l_q^*, \hat{\gamma} - \lambda/2) \hat{c}}{\hat{t}_R^2 CV_{\min} + X(l_q^*, \hat{\gamma}) \hat{c}} \quad (4.2)$$

where  $X(l_q^*, \lambda/2) = \sum_{k=1}^{N-l_q^*} x_{(k)}^{\lambda/2}$ ,  $X(l_q^*, \hat{\gamma} - \lambda/2) = \sum_{k=1}^{N-l_q^*} x_{(k)}^{\hat{\gamma}-\lambda/2}$ , and  $\hat{t}_R$  may be estimated from past survey data corresponding to the period of the auxiliary information. The total new minimal sample size required to meet the precision constraint is then given by

$$n_q^* = n_{aq}^* + n_{bq}^* = l_q^*(n_q) + n_{bq}^*. \quad (4.3)$$

### 4.2 Combined Sample Redesign Methodology

Next, note that the solution to the Transfer Algorithm  $l^*$  depends on the current total sample size:  $l_q^*(\lambda) \equiv l_q^*(\lambda, n_q)$ . Once the new minimal sample size  $n_q^*$  is determined, the existing partitioning  $\{U_{aq}^*(l_q^*), U_{bq}^*(l_q^*)\}$  which was optimal at  $n_q$  is no longer optimal at the new minimal sample size  $n_q^*$  because  $l^*(\lambda, n_q^*) \neq l^*(\lambda, n_q)$  if  $n_q^* \neq n_q$ . Therefore, letting  $n_{q+1} = n_q^*$ , a new population partitioning from the Transfer Algorithm based on  $l_{q+1}^*(\lambda, n_{q+1})$ , given by  $\{U_{a,q+1}^*(l_{q+1}^*), U_{b,q+1}^*(l_{q+1}^*)\}$ , is required to optimize the construction of the take-all and sampled sub-populations. Next, applying (4.2) over  $U_{b,q+1}^*(l_{q+1}^*)$  gives a new minimal sample size  $n_{q+1}^* = l_{q+1}^*(n_{q+1}) + n_{b,q+1}^*$  required to achieve the desired precision  $CV_{\min}$ . Proceeding in this fashion, the combined scheme produces a sequence of population partitionings, sample sizes, and sample allocations

$$(l^*(\lambda, n_q), (n_{aq} = l_q^*, n_{bq} = n_q - l_q^*)),$$

$$(N_{aq}^* = l_q^*, N_{bq}^* = N - l_q^*), (n_{aq}^* = l_q^*, n_{bq}^*), q = 0, 1, \dots \quad (4.4)$$

with  $n_{q+1} = n_q^* = n_{aq}^* + n_{bq}^*$  and the initial value  $n_0$  (current survey sample size). The combined procedure is repeated until further reductions in the minimal sample size can no longer be achieved. This leads to the stopping rule

$$q^* = \min_q \{q : n_{q+1}^* - n_q^* \geq 0\}. \quad (4.5)$$



The optimality of the combined procedure can be established using Theorem 2 and is omitted here due to space (see Pandher 1995). The main result is that the combined procedure converges to a globally optimal solution along the path defined by (4.4) to a point where further reductions in the sample size are not possible (by reconstructing  $U_a^*$  and  $U_b^*$ ) given the imposed precision constraint.

## 5. APPLICATION

The combined sample design procedure described above is now applied to the redesign of the Local Government Finance Survey in the province of Ontario. The survey response  $y$  in this application is the actual expenditures reported for sampled local government units for Ontario in 1989. The actual estimates are prepared 30 months after the end of the survey year from financial statements submitted by the local government units to the Department of Municipal Affairs (provincial). Population counts for the local government units from the nearest census (1991) are used as the auxiliary variable  $x$ . The population of local-level municipalities for Ontario consists of a total of 793 units of which a sample of 108 units is currently taken.

The results of applying the combined methodology to Ontario LGFS data are reported in Table 1. The level of desired precision  $CV_{\min}$  was set at 2% for the total regression estimator  $\hat{t}_R = t_a + \hat{t}_{Rb}$ . Using the methods of Pandher (1995), the best value for the heteroscedasticity parameter  $\gamma$  in Ontario was determined to be  $\hat{\gamma} = 2$ ; the corresponding proportionality constant was estimated to be  $\hat{c} = .0825$ . The near optimal sample design defined by  $\lambda = \hat{\gamma} (p(s; \hat{\gamma}))$  was used.

**Table 1**  
Application of Combined Methodology to LGF Survey Data  
(Ontario, 1989)

Iteration ( $q$ )	$n_q$	$l_q(\lambda, n_q)$	$n_{aq}^*$	$n_{bq}^*$	$n_q^*$
0	108	39	39	18	57
1	57	16	16	34	50
2	50	12	12	38	50

For Ontario the combined scheme stopped at iteration  $q^* = 2$ . The globally optimal population partitioning between the take-all and sampled groups is  $N_a^* = 16$  and  $N_b^* = 777$ . The new minimal total sample size is  $n^* = 50$  with allocations  $n_a^* = 16$  and  $n_b^* = 34$ . A total sample size reduction of  $n_0 - n_2^* = 108 - 50 = 58$  is achieved at the desired CV of 2% for the regression estimator  $\hat{t}_R = t_a + \hat{t}_{Rb}$ .

## 6. CONCLUDING REMARKS

This paper provides a comprehensive methodology for identifying and implementing an efficient sample design for recurrent surveys of skewed populations. The combined

procedure integrates the solution to the following three problems: i) identifying an efficient sample selection scheme, ii) constructing an efficient demarcation between the take-all and sampled population groups at a given sample size, and iii) determining the minimal sample size required to meet the precision constraint(s).

The equivalence result to the Transfer Algorithm (Pandher 1995) was used to create the take-all and sampled groups. The first two components were then combined with a sample size determination step through an iterative procedure. Under the stopping rule, the combined iterative procedure converges to a globally minimal sample size and optimal population partitioning. Results from the application of the proposed sample redesign methodology to the Local Government Survey in Ontario were reported. A 52% reduction in the total sample size was achieved for the regression estimator of the total ( $\hat{t}_R = t_a + \hat{t}_{Rb}$ ) at the desired precision of  $CV = 2\%$ .

## ACKNOWLEDGEMENTS

The author would like to acknowledge the support of Public Institutions Division for sponsoring this work and thank M.P. Singh, H. Mantel, M.S. Kovacevic, S. Wu and the referees for their valuable comments on earlier drafts of the paper.

## APPENDIX

A brief sketch of the development behind Theorem 2 (Equivalence Result) is given here; for technical details see Pandher (1995). The same paper also establishes the desirable mathematical properties of the Transfer Algorithm such as existence and optimality of solution as well as the optimality of the combined procedure.

Using the expression for the variance of  $V^{(l)}(\hat{t}_{Rb}; \cdot)$  given in (3.1), the difference  $V^{(l+1)} - V^{(l)}$  may be expressed as

$$\hat{V}^{(l+1)} - \hat{V}^{(l)} = c \frac{A(l) B(l)}{(n-l)(n-l-1)} \quad (A.1)$$

where

$$A(l) = \sum_{j=1}^{N-l} x_{(j)}^{\lambda/2} - (n-l) x_{(N-l)}^{\lambda/2}$$

and

$$B(l) = \sum_{k=1}^{N-l} x_{(k)}^{\gamma-\lambda/2} - (n-l) x_{(N-l)}^{\gamma-\lambda/2}.$$

The condition  $B(l) < 0$  may also be expressed as  $n-l > R(l; \gamma - \lambda/2)$  where  $R(l; \alpha) = \sum_{k=1}^{N-l} x_{(k)}^{\alpha} / x_{(N-l)}^{\alpha}$ . Similarly, the condition  $A(l) > 0$  corresponds to  $n-l < R(l; \lambda/2)$ . All possible states of the system defined by the Transfer Algorithm are summarized in Table A.1.

**Table A.1**  
Outcomes for  $V^{(l+1)} - V^{(l)} < 0$  and  $V^{(l+1)} - V^{(l)} \geq 0$   
in Terms of  $n^{(l)} = n - l$

Behaviour of A and B	$V^{(l+1)} - V^{(l)} < 0$
	Condition on $n^{(l)} = n - l$
$A(l) > 0$ $B(l) < 0$	$R(l; \gamma - \lambda/2) < n - l < R(l; \lambda/2)$ (T.1)
$A(l) < 0$ $B(l) > 0$	$R(l; \lambda/2) < n - l < R(l; \gamma - \lambda/2)$ (T.3)
	$V^{(l+1)} - V^{(l)} \geq 0$
	Condition on $n^{(l)} = n - l$
$A(l) > 0$ $B(l) \geq 0$	$n - l \leq \min\{R(l; \lambda/2), R(l; \gamma - \lambda/2)\}$ (T.2)
$A(l) \leq 0$ $B(l) \leq 0$	$n - l \geq \max\{R(l; \lambda/2), R(l; \gamma - \lambda/2)\}$ (T.4)

The first column describes the behaviour of  $A(l)$  and  $B(l)$  leading to the outcome  $V^{(l+1)} - V^{(l)} < 0$  and  $V^{(l+1)} - V^{(l)} \geq 0$ , respectively. The second column describes the equivalent condition in terms of  $n^{(l)} = n - l$ ,  $R(l; \gamma - \lambda/2)$ , and  $R(l; \lambda/2)$  corresponding to  $V^{(l)} - V^{(l-1)} < 0$  and  $V^{(l)} - V^{(l-1)} \geq 0$ , respectively. An important condition required for the solution to the Transfer Algorithm  $l^*(\lambda)$  is that  $\pi_{(N-l)}(\lambda) < 1$  hold. It is easy to verify that  $\pi_{(N-l)}(\lambda) < 1 \Leftrightarrow A(l) > 0$ . In terms of the description for the Transfer Algorithm given in Table A.1, this condition means that the solution can occur only when both  $A(l) > 0$  and  $B(l) \geq 0$  or, equivalently, when  $n - l$  satisfies condition (T.2).

Table A.1 completely enumerates all possible states of the system defined by the Transfer Algorithm. The correspondence between the internal cell quantities (computable directly from the auxiliary data and estimated parameters) and the margins ( $A(l)$ ,  $B(l)$ ,  $V^{(l+1)} - V^{(l)}$ ) represents a tautology

which leads directly to Theorem 2 (Equivalence Result). The behaviour of the system described in the table also depends on the sample design  $p(s; \lambda)$  employed. The three relevant cases are:

- $0 \leq \lambda < \gamma \Rightarrow [R(l; \gamma - \lambda/2) < R(l; \lambda/2)]$ ,
- $\lambda = \gamma \Rightarrow [R(l; \gamma - \lambda/2) = R(l; \lambda/2)]$ , and
- $\gamma < \lambda \Rightarrow [R(l; \gamma - \lambda/2) > R(l; \lambda/2)]$ .

In case a) the system starts ( $l = 0$ ) in state (T.4), moves to (T.1) and then finally rests in state (T.2); state (T.3) is infeasible here. The solution to the Transfer Algorithm  $l^*(\lambda)$  is given by the smallest  $l$  leading the system to move into state (T.2). In case b), the system starts in state (T.4) and moves to (T.2); (T.1) and (T.3) do not apply. Finally, in case c), the transition path is from (T.4) to (T.3) to (T.2); here (T.1) is invalid.

## REFERENCES

- GLASSER, G.J. (1962). On the complete coverage of large units in a statistical study. *Review of the International Statistical Institute*, 30, 28-32.
- GODAMBE, V.P., and JOSHI, V.M. (1965). Admissibility and Bayes estimation in sampling finite populations. *Annals of Mathematical Statistics*, 36, 1702-1722.
- HIDIROGLOU, M.A., and SRINATH, K.P. (1993). Problems associated with designing subannual business surveys. *Journal of Business and Economic Statistics*, 11, 397-405.
- LAVALLÉE, P., and HIDIROGLOU, M.A. (1988). On the stratification of skewed populations. *Survey Methodology*, 14, 33-43.
- PANDHER, G.S. (1995). Surveys of skewed populations: optimal sample redesign under the generalized regression estimator with applications to the Local Government Finance Survey. Methodology Branch Working Paper: HSMD-95-006, Statistics Canada.
- SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.



## ACKNOWLEDGEMENTS

*Survey Methodology* wishes to thank the following persons who have served as referees during 1996. An asterisk indicates that the person served more than once.

- Y. Amamiya, *Iowa State University*  
M. Bankier, *Statistics Canada*  
\* D.R. Bellhouse, *University of Western Ontario*  
T.R. Belin, *University of California - Los Angeles*  
W. Bell, *U.S. Bureau of the Census*  
\* D.A. Binder, *Statistics Canada*  
F.J. Breidt, *Iowa State University*  
K. Campbell, *Los Alamos National Laboratory*  
C.M. Cassel, *VIKON, Statistics Sweden*  
G. Chen, *PNN Laboratory*  
G. Chen, *University of Regina*  
J. Chen, *University of Waterloo*  
C. Cordy, *Oregon State University*  
C. D. Cowan, *Federal Deposit Insurance Corporation*  
B.G. Cox, *Mathematica Policy Research, Inc.*  
C. S. Davis, *University of Iowa*  
\* J. Denis, *Statistics Canada*  
J.-C. Deville, *INSEE*  
A.H. Dorfman, *U.S. Bureau of Labor Statistics*  
J.D. Drew, *Statistics Canada*  
J.-J. Dreesbeke, *Université Libre de Bruxelles*  
P. Duchesne, *Université de Montréal*  
J.L. Eltinge, *Texas A&M University*  
L. Ernst, *U.S. Bureau of Labor Statistics*  
J.T. Fagan, *U.S. Bureau of the Census*  
W.A. Fuller, *Iowa State University*  
\* J. Gambino, *Statistics Canada*  
Y. Gervais, *Statistics Canada*  
R.M. Groves, *University of Maryland*  
\* M.A. Hidioglou, *Statistics Canada*  
D. Holt, *Central Statistical Office, U.K.*  
E. Hoy, *U.S. Bureau of the Census*  
E. Johnson, *Educational Testing Service*  
C. Julien, *Statistics Canada*  
G. Kalton, *Westat, Inc.*  
\* P.S. Kott, *National Agricultural Statistical Service*  
M. Kovacevic, *Statistics Canada*  
R. Lachapelle, *Statistics Canada*  
M. Latouche, *Statistics Canada*  
P. Lavallée, *Statistics Canada*  
S. Linacre, *Australian Bureau of Statistics*  
\* D. Malec, *National Center for Health Statistics*  
\* H. Mantel, *Statistics Canada*  
H. Mariotte, *INSEE*  
A. Mason, *East-West Center*  
N. Mathiowetz, *University of Maryland*  
S.M. Miller, *U.S. Bureau of Labor Statistics*  
P.L. do Nascimento Silva, *University of Southampton*  
L. Norberg, *Statistics Sweden*  
D. Pfeffermann, *Hebrew University*  
J. Qian, *National Opinion Research Center*  
R. Raby, *Statistics Canada*  
T.E. Raghunathan, *University of Michigan*  
\* J.N.K. Rao, *Carleton University*  
\* L.-P. Rivest, *Université Laval*  
G. Roberts, *Statistics Canada*  
K. Rust, *Westat, Inc.*  
I. Sande, *Bell Communications Research, U.S.A.*  
\* C.-E. Särndal, *Université de Montréal*  
\* W.L. Schaible, *U.S. Bureau of Labor Statistics*  
N. Schenker, *University of California - Los Angeles*  
F.J. Scheuren, *George Washington University*  
J. Sedransk, *Case Western Reserve University*  
J.P. Shaffer, *Educational Testing Service*  
G.M. Shapiro, *Westat Inc.*  
\* A.C. Singh, *Statistics Canada*  
\* M.P. Singh, *Statistics Canada*  
\* R. Sitter, *Simon Fraser University*  
C.J. Skinner, *University of Southampton*  
E.A. Stasny, *Ohio State University*  
\* D. Stukel, *Statistics Canada*  
\* R. Thomas, *Carleton University*  
M. Thompson, *University of Waterloo*  
Y. Tillé, *École nationale de statistique et de l'analyse de l'information*  
D. Tiller, *U.S. Bureau of Labor Statistics*  
R. Tourangeau, *National Opinion Research Center*  
R. Valliant, *U.S. Bureau of Labor Statistics*  
R.B.P. Verma, *Statistics Canada*  
V.K. Verma, *University of Essex*  
P.J. Waite, *U.S. Bureau of the Census*  
J. Waksberg, *Westat, Inc.*  
W.E. Winkler, *U.S. Bureau of the Census*  
K.M. Wolter, *National Opinion Research Center*  
S. Wu, *Statistics Canada*  
E. Zanutto, *Harvard University*  
\* A. Zaslavsky, *Harvard University*

Acknowledgements are also due to those who assisted during the production of the 1996 issues: S. Beauchamp and L. Durocher (Composition Unit) and L. Perreault (Official Languages and Translation Division). Finally we wish to acknowledge S. DiLoreto, S.F. Bertrand, C. Larabie and D. Lemire of Household Survey Methods Division, for their support with coordination, typing and copy editing.

# The Statistician

JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES D

---

## CONTENTS

VOLUME 45, No. 2, 1996

---

- Trend growth in post-1850 British economic history: the Kalman filter and historical judgement  
*M. Ball and A. Wood*
- Modelling trends in economic history  
*T.C. Mills and N.F.R. Crafts*
- Some comments on modelling economic trends  
*M. Ball and A. Wood*
- Exploring survey non-response: the effect of attrition on a follow-up of the 1984-85 health and life style survey  
*R. Gray, P. Campanelli, K. Deepchand and P. Prescott-Clarke*
- A graphical approach to identification of dependent failures  
*L. Walls*
- Infinite estimates with fractional factorial experiments  
*A.N. Pettitt*
- On the use of the Leslie matrix in the analysis of prevalence data for general epidemic processes  
*F.W.O. Saporu*
- Estimating the proportion of lymphoblastoid cells affected by exposure to ethylene oxide through micronuclei counts  
*J.K. Lindsey and C. Laurent*
- Retrospective power surveys  
*R. Gillett*
- Modification of Waterton's controlled selection method  
*S.S.A. Ghazali*
- Focus on sport*
- Lower bounds for athletic performance  
*D.C. Blest*
- A comparison of leg before wicket rates between Australians and their visiting teams for test cricket series played in Australia, 1977-94  
*S.M. Crowe and J.M. Middeldorp*
- 

*Correspondence*

*Book reviews*

THE ROYAL STATISTICAL SOCIETY, 12 ERROL ST, LONDON, EC1Y 8LX, UK



## CONTENTS

## TABLE DES MATIÈRES

## Volume 24, No. 3, September/septembre 1996

- Deli LI  
On moments of the supremum of normed weighted averages
- André Robert DABROWSKI and David MacDONALD  
An application of the Bernoulli part to local limit theorems for moving averages on stationary sequences
- Paul GUSTAFSON  
The effect of mixing-distribution misspecification in conjugate mixture models
- Hyun Suk LEE  
Analysis of overdispersed paired count data
- James H. ALBERT  
Bayesian selection of log-linear models
- Paul CABILIO and Joe MASARO  
A simple test of symmetry about an unknown median
- Gemai CHEN  
EDF tests of goodness-of-fit for transform-both-sides models
- D.N. SHAH and P.A. PATEL  
Asymptotic properties of a generalized regression-type predictor of a finite population variance in probability sampling
- Bent JØRGENSEN, Søren LUNDBYE-CHRISTENSEN, Xue-Kun SONG and Li SUN  
State-space models for multivariate longitudinal data of mixed types

## Volume 24, No. 4, December/décembre 1996

- Miklós CSÖRGÖ and Hao YU  
Weak approximations for quantile processes of stationary sequences
- Karen Y. FUNG, D. KREWSKI and R.T. SMYTHE  
A comparison of tests for trend with historical controls in carcinogen bioassay
- Lise MANCHESTER and Wade BLANCHARD  
When is a curve an outlier? An account of a tricky problem
- Philippe C. BESSE et Hervé CARDOT  
Approximation spline de la prévision d'un processus fonctionnel autorégressif d'ordre 1
- H. WONG and W.K. LI  
Distribution of the cross-correlations of squared residuals in ARIMA models
- A.L. RUKHIN  
Linear statistics in change-point estimation and their asymptotic behavior
- Ellen MAKI and Philip McDUNNOUGH  
The role of probability generating functions for estimation in incompletely observed random walks
- Paul GUSTAFSON  
Model influence functions based on mixtures
- Majid MOJIRSHEIBANI and Robert TIBSHIRANI  
Some results on bootstrap prediction intervals
- Constantinos GOUTIS and Rex F. GALBRAITH  
A parametric model for heterogeneity in paired Poisson counts
- Nicolas W. HENGARTNER and Oliver B. LINTON  
Nonparametric regression estimation at design poles and zeros

# JOURNAL OF OFFICIAL STATISTICS

## An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

### Contents

#### Volume 12, Number 1, 1996

Robust Case-Weighting for Multipurpose Establishment Surveys <i>R.L. Chambers</i> .....	3
Maximizing the Overlap of Sample Units for Two Designs with Simultaneous Selection <i>Lawrence R. Ernst</i> .....	33
Estimates of National Hospital Use from Administrative Data and Personal Interviews <i>Sally Stearns, Mary Grace Kovar, Kevin Hayes, and Gary Koch</i> .....	47
Contact-Level Influences on Cooperation in Face-to-Face Surveys <i>Robert M. Groves and Mick P. Couper</i> .....	63
A Bayesian Approach to Designing U.S. Census Sampling for Reapportionment <i>Joseph B. Kadane</i> .....	85
Comments on "Designing Census Sampling for Apportionment": Sample Design for a Multi-Purpose Census <i>Alan M. Zaslavsky</i> .....	95
Comments on "Designing Census Sampling for Apportionment" <i>Mary M. Mulry</i> .....	101
Rejoinder <i>Joseph B. Kadane</i> .....	105
Book Review .....	107

#### Volume 12, Number 2, 1996

Why Innovation is Difficult in Government Surveys <i>Don A. Dillman</i> .....	113
Comment <i>Barbara A. Bailer</i> .....	125
Comment <i>Jelke Bethlehem</i> .....	129
Comment <i>David A. Binder</i> .....	133
Comment <i>Barbara Everitt Bryant</i> .....	137
Comment <i>Cynthia Z.F. Clark</i> .....	141
Comment <i>Michael Colledge</i> .....	145
Comment <i>I.P. Fellegi</i> .....	151
Comment <i>Stephen E. Fienberg and Judith M. Tanur</i> .....	157
Comment <i>Eivind Hoffman</i> .....	161
Comment <i>C.L. Kincannon</i> .....	165
Comment <i>Susan M. Miskura</i> .....	169
Comment <i>Thomas J. Plewes</i> .....	171
Comment <i>Wesley L. Schaible</i> .....	175
Comment <i>Robert D. Tortora</i> .....	179
Comment <i>Dennis Trewin</i> .....	185
Rejoinder <i>Don A. Dillman</i> .....	191
A Comparison of Ten Methods for Multilateral International Price and Volume Comparison <i>Bert Balk</i> .....	199
In Other Journals .....	223

All inquiries about submissions and subscriptions should be directed to the Chief Editor:  
Lars Lyberg, R&D Department, Statistics Sweden, S - 115 81 Stockholm, Sweden.



# GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue (Vol. 19, No. 1 and onward) of *Survey Methodology* as a guide and note particularly the following points:

## 1. Layout

- 1.1 Manuscripts should be typed on white bond paper of standard size ( $8\frac{1}{2} \times 11$  inch), one side only, entirely double spaced with margins of at least  $1\frac{1}{2}$  inches on all sides.
- 1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
- 1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

## 2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

## 3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as “exp(·)” and “log(·)”, etc.
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (e.g., w,  $\omega$ ; o, O; 0; l, 1).
- 3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

## 4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
- 4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

## 5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

# DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, prière d'examiner un numéro récent de *Techniques d'enquête* (à partir du vol. 19, n° 1) et de noter les points suivants:

## 1. Présentation

- 1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8½ par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.
- 1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.
- 1.3 Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
- 1.4 Les remerciements doivent paraître à la fin du texte.
- 1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.

## 2. Résumé

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

## 3. Rédaction

- 3.1 Éviter les notes au bas des pages, les abréviations et les sigles.
- 3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp(-) et log(·) etc.
- 3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
- 3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.
- 3.5 Distinguer clairement les caractères ambigus (comme w, ω; o, O, 0; l, 1).
- 3.6 Les caractères italiques sont utilisés pour faire ressortir des mots. Indiquer ce qui doit être imprimé en italique en le soulignant dans le texte.

## 4. Figures et tableaux

- 4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
- 4.2 Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois).

## 5. Bibliographie

- 5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence.  
Exemple: Cochran (1977, p. 164).
- 5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.



JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

## Contents

## Volume 12, Number 1, 1996

Robust Case-Weighting for Multipurpose Establishment Surveys

*R.L. Chambers* . . . . . 3

Maximizing the Overlap of Sample Units for Two Designs with Simultaneous Selection

*Lawrence R. Emst* . . . . . 33

Estimates of National Hospital Use from Administrative Data and Personal Interviews

*Sally Stearns, Mary Grace Kovar, Kevin Hayes, and Gary Koch* . . . . . 47

Contact-Level Influences on Cooperation in Face-to-Face Surveys

*Robert M. Groves and Mick P. Couper* . . . . . 63

A Bayesian Approach to Designing U.S. Census Sampling for Reapportionment

*Joseph B. Kadane* . . . . . 85

Comments on "Designing Census Sampling for Apportionment": Sample Design for a Multi-Purpose Census

*Alan M. Zaslavsky* . . . . . 95

Comments on "Designing Census Sampling for Apportionment"

*Mary M. Mulry* . . . . . 101

Rejoinder

*Joseph B. Kadane* . . . . . 105

Book Review . . . . . 107

## Volume 12, Number 2, 1996

Why Innovation is Difficult in Government Surveys

*Don A. Dillman* . . . . . 113

Comment

*Barbara A. Bailar* . . . . . 125

Comment

*Jelke Beilharz* . . . . . 129

Comment

*David A. Binder* . . . . . 133

Comment

*Barbara Everitt Bryant* . . . . . 137

Comment

*Cynthia Z.F. Clark* . . . . . 141

Comment

*Michael Colledge* . . . . . 145

Comment

*I.P. Fellegi* . . . . . 151

Comment

*Stephen E. Fienberg and Judith M. Tanur* . . . . . 157

Comment

*Eivind Hoffman* . . . . . 161

Comment

*C.T. Kincannon* . . . . . 165

Comment

*Susan M. Miskura* . . . . . 169

Comment

*Thomas J. Plewes* . . . . . 171

Comment

*Wesley L. Schabile* . . . . . 175

Comment

*Robert D. Tortora* . . . . . 179

Comment

*Dennis Trewin* . . . . . 185

Rejoinder

*Don A. Dillman* . . . . . 191

A Comparison of Ten Methods for Multilateral International Price and Volume Comparison

*Bert Balk* . . . . . 199

In Other Journals

223

Volume 24, No. 3, September/septembre 1996

Deji LI  
On moments of the supremum of normed weighted averages  
André Robert DABROWSKI and David MacDONALD  
An application of the Bernoulli part to local limit theorems for moving averages on stationary sequences

Paul GUSTAFSON  
The effect of mixing-distribution misspecification in conjugate mixture models

Hyun Suk LEE  
Analysis of overdispersed paired count data

James H. ALBERT

Bayesian selection of log-linear models

Paul CABILLO and Joe MASARO

A simple test of symmetry about an unknown median

Gemai CHEN

EDF tests of goodness-of-fit for transform-both-sides models

D.N. SHAH and P.A. PATEL

Asymptotic properties of a generalized regression-type predictor of a finite population variance in probability sampling

Bent JØRGENSEN, Søren LUNDBYE-CHRISTENSEN, Xue-Kun SONG and Li SUN

State-space models for multivariate longitudinal data of mixed types

Volume 24, No. 4, December/décembre 1996

Miklós CSÖRGÖ and Hao YU

Weak approximations for quantile processes of stationary sequences

Karen Y. FUNG, D. KREWISKI and R.T. SMYTHE

A comparison of tests for trend with historical controls in carcinogen bioassay

Lise MANCHESTER and Wade BLANCHARD

When is a curve an outlier? An account of a tricky problem

Philippe C. BESSÉ et Hervé CARDOT

Approximation spline de la prévision d'un processus fonctionnel autorégressif d'ordre 1

H. WONG and W.K. LI

Distribution of the cross-correlations of squared residuals in ARIMA models

A.L. RUKHIN

Linear statistics in change-point estimation and their asymptotic behavior

Ellen MAKI and Philip McDUNNOUGH

The role of probability generating functions for estimation in incompletely observed random walks

Paul GUSTAFSON

Model influence functions based on mixtures

Majid MOJIRSHHEIBANI and Robert TIBSHIRANI

Some results on bootstrap prediction intervals

Constantinos GOUTIS and Rex F. GALBRAITH

A parametric model for heterogeneity in paired Poisson counts

Nicolas W. HENGARTNER and Oliver B. LINTON

Nonparametric regression estimation at design poles and zeros



# The Statistician

JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES D

CONTENTS VOLUME 45, No. 2, 1996

Trend growth in post-1850 British economic history: the Kalman filter and historical judgement  
*M. Ball and A. Wood*

Modelling trends in economic history  
*T.C. Mills and N.F.R. Crafts*

Some comments on modelling economic trends  
*M. Ball and A. Wood*

Exploring survey non-response: the effect of attrition on a follow-up of the 1984-85 health and life style survey  
*R. Gray, P. Campanelli, K. Deepchand and P. Prescott-Clarke*

A graphical approach to identification of dependent failures  
*L. Wallis*

Infinite estimates with fractional factorial experiments  
*A.N. Pettitt*

On the use of the Leslie matrix in the analysis of prevalence data for general epidemic processes  
*F.W.O. Sapornu*

Estimating the proportion of lymphoblastoid cells affected by exposure to ethylene oxide through micronuclei counts  
*J.K. Lindsey and C. Laurent*

Retrospective power surveys  
*R. Gillett*

Modification of Waterton's controlled selection method  
*S.S.A. Ghazali*

*Focus on sport*  
Lower bounds for athletic performance  
*D.C. Blest*

A comparison of leg before wicket rates between Australians and their visiting teams for test cricket series played in Australia, 1977-94  
*S.M. Crowe and J.M. Middelkamp*

Correspondence  
Book reviews

THE ROYAL STATISTICAL SOCIETY, 12 ERROL ST, LONDON, EC1Y 8LX, UK

## REMERCIEMENTS

*Techniques d'enquête* désire remercier les personnes suivantes, qui ont accepté de faire la critique d'un article durant l'année 1996. Un astérisque indique que la personne a participé plus d'une fois.

Y. Amamiya, *Iowa State University*  
 M. Bankier, *Statistique Canada*  
 \* D.R. Bellhouse, *University of Western Ontario*  
 T.R. Bellin, *University of California - Los Angeles*  
 W. Bell, *U.S. Bureau of the Census*  
 \* D.A. Binder, *Statistique Canada*  
 F.J. Breidt, *Iowa State University*  
 K. Campbell, *Los Alamos National Laboratory*  
 C.M. Cassel, *VIKON, Statistics Sweden*  
 G. Chen, *PNN Laboratory*  
 G. Chen, *University of Regina*  
 J. Chen, *University of Waterloo*  
 C. Cordy, *Oregon State University*  
 C. D. Cowan, *Federal Deposit Insurance Corporation*  
 B.G. Cox, *Mathematica Policy Research, Inc.*  
 C. S. Davis, *University of Iowa*  
 \* J. Denis, *Statistique Canada*  
 J.-C. Deville, *INSEE*  
 A.H. Dorfman, *U.S. Bureau of Labor Statistics*  
 J.D. Drew, *Statistique Canada*  
 J.-J. Droesbeke, *Université Libre de Bruxelles*  
 P. Duchesne, *Université de Montréal*  
 J.L. Eltinge, *Texas A&M University*  
 L. Ernst, *U.S. Bureau of Labor Statistics*  
 J.T. Fagan, *U.S. Bureau of the Census*  
 W.A. Fuller, *Iowa State University*  
 \* J. Gambino, *Statistique Canada*  
 Y. Gervais, *Statistique Canada*  
 R.M. Groves, *University of Maryland*  
 \* M.A. Hidiroglou, *Statistique Canada*  
 D. Holt, *Central Statistical Office, U.K.*  
 E. Hoy, *U.S. Bureau of the Census*  
 E. Johnson, *Educational Testing Service*  
 C. Julien, *Statistique Canada*  
 G. Kalton, *Westat, Inc.*  
 \* P.S. Kott, *National Agricultural Statistical Service*  
 M. Kovacevic, *Statistique Canada*  
 R. Lachapelle, *Statistique Canada*  
 M. Latouche, *Statistique Canada*  
 P. Lavallée, *Statistique Canada*  
 S. Linacre, *Australian Bureau of Statistics*  
 \* D. Malec, *National Center for Health Statistics*  
 \* H. Mantel, *Statistique Canada*  
 H. Mariotte, *INSEE*

On remercie également ceux qui ont contribué à la production des numéros de la revue pour 1996: S. Beauchamp et L. Durocher (Unité de composition) et L. Perreault (Division des langues officielles et traduction). Finalement on désire exprimer notre reconnaissance à S. DiLoreto, S.F. Bertrand, C. Larabie et D. Lémire de la Division des méthodes d'enquêtes des ménages, pour leur apport à la coordination, la dactylographie et la rédaction.

A. Mason, *East-West Center*  
 N. Mathiowetz, *University of Maryland*  
 S.M. Miller, *U.S. Bureau of Labor Statistics*  
 P.L. do Nascimento Silva, *University of Southampton*  
 L. Norberg, *Statistics Sweden*  
 D. Pfeffermann, *Hebrew University*  
 J. Qian, *National Opinion Research Center*  
 R. Raby, *Statistique Canada*  
 T.E. Raghunathan, *University of Michigan*  
 \* J.N.K. Rao, *Carleton University*  
 \* L.-P. Rivest, *Université Laval*  
 G. Roberts, *Statistique Canada*  
 K. Rust, *Westat, Inc.*  
 I. Sande, *Bell Communications Research, U.S.A.*  
 \* C.-E. Särndal, *Université de Montréal*  
 \* W.L. Schaible, *U.S. Bureau of Labor Statistics*  
 N. Schenker, *University of California - Los Angeles*  
 F.J. Scheuren, *George Washington University*  
 J. Sedransk, *Case Western Reserve University*  
 J.P. Shaffer, *Educational Testing Service*  
 G.M. Shapiro, *Westat Inc.*  
 \* A.C. Singh, *Statistique Canada*  
 \* M.P. Singh, *Statistique Canada*  
 \* R. Sitter, *Simon Fraser University*  
 C.J. Skinner, *University of Southampton*  
 E.A. Stasny, *Ohio State University*  
 \* D. Stukel, *Statistique Canada*  
 \* R. Thomas, *Carleton University*  
 M. Thompson, *University of Waterloo*  
 Y. Tillé, *École nationale de statistique et de l'analyse de l'information*  
 D. Tiller, *U.S. Bureau of Labor Statistics*  
 R. Tourangeau, *National Opinion Research Center*  
 R. Valliant, *U.S. Bureau of Labor Statistics*  
 R.B.P. Verma, *Statistique Canada*  
 V.K. Verma, *University of Essex*  
 P.J. Waite, *U.S. Bureau of the Census*  
 J. Waksberg, *Westat, Inc.*  
 W.E. Winkler, *U.S. Bureau of the Census*  
 K.M. Wolter, *National Opinion Research Center*  
 S. Wu, *Statistique Canada*  
 E. Zanutto, *Harvard University*  
 \* A. Zaslavsky, *Harvard University*





## BIBLIOGRAPHIE

- celles des marges ( $A(l)$ ,  $B(l)$ ,  $V^{(l+1)} - V^{(l)}$ ) représente une tauologie qui mène directement au théorème 2 (solution équivalente). Le comportement du système décrit dans le tableau dépend également du plan d'échantillonnage  $p(s; \lambda)$  choisi. Les trois cas pertinents sont:
- a)  $0 \leq \lambda < \gamma \Rightarrow [R(l; \gamma - \lambda/2) < R(l; \lambda/2)]$ ,  
 b)  $\lambda = \gamma \Rightarrow [R(l; \gamma - \lambda/2) = R(l; \lambda/2)]$ , et  
 c)  $\gamma < \lambda \Rightarrow [R(l; \gamma - \lambda/2) > R(l; \lambda/2)]$ .
- Dans le cas a), le système débute ( $l = 0$ ) à l'état (T.4), passe à (T.1) puis finalement s'arrête à l'état (T.2); l'état (T.3) est impossible ici. La solution de l'algorithme de transfert  $l^*(\lambda)$  est donnée par la plus petite valeur de  $l$  amenant le système à passer à l'état T.2. Dans le cas b), le système débute à l'état (T.4) et passe à l'état (T.2); (T.1) et (T.3) ne sont pas applicables. Enfin, dans le cas c), le cheminement se fait de (T.4) à (T.3), puis à (T.2); ici, (T.1) n'est pas valide.
- GLASSER, G.J. (1962). On the complete coverage of large units in a statistical study. *Revue de l'Institut International de Statistique*, 30, 28-32.
- GODAMBE, V.P., et JOSHI, V.M. (1965). Admissibility and Bayes estimation in sampling finite populations. *Annals of Mathematical Statistics*, 36, 1702-1722.
- HIDIROGLOU, M.A., et SRINATH, K.P. (1993). Problems associated with designing subannual business surveys. *Journal of Business and Economic Statistics*, 11, 397-405.
- LAVALLÉE, P., et HIDIROGLOU, M.A. (1988). Sur la stratification de populations asymétriques. *Techniques d'enquête*, 14, 35-45.
- PANDHER, G.S. (1995). Surveys of skewed populations: optimal sample redesign under the generalized regression estimator with applications to the Local Government Finance Survey. Documents de travail de la Direction de la méthodologie: HSM-D-95-006, Statistique Canada.
- SÄRNDAAL, C.-E., SWENSSON, B., et WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.



6. CONCLUSIONS

Le présent article décrit une méthode générale permettant de déterminer et d'appliquer un plan d'échantillonnage efficace pour les enquêtes récurrentes effectuées sur des populations à distribution asymétriques. La méthode combinée fournit et intègre les solutions de trois problèmes, soit: i) définir un plan d'échantillonnage efficace, ii) construire une partition efficace de la population en une sous-population à tirage complet et une sous-population échantillon pour une taille donnée d'échantillon et iii) déterminer la taille minimale de l'échantillon nécessaire pour satisfaire la ou les contraintes de précision.

On a utilisé la solution équivalente proposée pour l'algorithme de transfert (Pandher 1995) pour créer les groupes à tirage complet et échantillon. Puis, on a combiné les deux premières composantes à une étape de détermination de la taille de l'échantillon grâce à une méthode itérative. Conformément à la règle d'arrêt, la méthode itérative combinée converge vers une taille globalement optimale d'échantillon et une partition globalement optimale de la population. Enfin, on a présenté les résultats de l'application de la méthode proposée de remaniement du plan d'échantillonnage à l'Enquête sur les finances des administrations locales effectuée en Ontario. L'application de la méthode a abouti à une réduction de 52% de la taille totale de l'échantillon pour l'estimateur de régression du total ( $t_R^* = t_a^* + t_{Rp}$ ) au degré de précision imposé, à savoir CV = 2%.

REMERCIEMENTS

L'auteur exprime sa reconnaissance à la Division des institutions publiques qui a accepté de parrainer les présents travaux et remercie M.P. Singh, H. Mantel, M.S. Kovacevic, S. Wu, et aux examinateurs pour les commentaires précieux qu'ils ont fait au sujet d'ébauches antérieures du présent article.

ANNEXE

Suit une brève description du développement du théorème 2 (solution équivalente). Pour les détails techniques, consulter Pandher (1995). Sont démontrées également dans le même article les propriétés mathématiques souhaitables de l'algorithme de transfert, dont l'existence d'une solution et l'optimalité de cette dernière, ainsi que l'optimalité de la méthode combinée.

En se servant de l'expression de la variance de  $V^{(t)}(f_{Rp}, \cdot)$  donnée en (3.1), on peut exprimer la différence  $V^{(t+1)} - V^{(t)}$  de la façon suivante:

$$V^{(t+1)} - V^{(t)} = c \frac{A(t) B(t)}{(n-l)(n-l-1)} \tag{A.1}$$

ou

$$A(t) = \sum_{l=1}^{f-1} x_{\lambda/2}^{(f)} - (n-l) x_{\lambda/2}^{(N-l)}$$

et

$$B(t) = \sum_{l=1}^{k-1} x_{\gamma-\lambda/2}^{(k)} - (n-l) x_{\gamma-\lambda/2}^{(N-l)}.$$

La condition  $B(t) < 0$  peut également être exprimée par  $n-l > R(t; \gamma - \lambda/2)$ , où  $R(t; \alpha) = \sum_{k=1}^{N-l} x_{\alpha}^{(k)} / x_{\alpha}^{(N-l)}$ . Pareillement, la condition  $A(t) > 0$  correspond à  $n-l < R(t; \lambda/2)$ . Tous les états possibles du système défini par l'algorithme de transfert sont résumés au tableau A.1.

Tableau A.1

Résultats pour $V^{(t+1)} - V^{(t)} < 0$ et $V^{(t+1)} - V^{(t)} \geq 0$ exprimés en fonction de $n^{(t)} = n - l$	
Comportement de A et de B	Condition imposée à $n^{(t)} = n - l$
$A(t) > 0$	$R(l; \gamma - \lambda/2) < n - l < R(l; \lambda/2)$ (T.1)
$A(t) < 0$	$R(l; \lambda/2) < n - l < R(l; \gamma - \lambda/2)$ (T.3)
$B(t) > 0$	$V^{(t+1)} - V^{(t)} \geq 0$ Condition imposée à $n^{(t)} = n - l$
$B(t) \leq 0$	$n - l \leq \min\{R(l; \lambda/2), R(l; \gamma - \lambda/2)\}$ (T.2)
$A(t) \leq 0$	$n - l \geq \max\{R(l; \lambda/2), R(l; \gamma - \lambda/2)\}$ (T.4)

Les première et troisième colonnes décrivent le comportement de  $A(t)$  et  $B(t)$  qui mène aux résultats  $V^{(t+1)} - V^{(t)} < 0$  et  $V^{(t+1)} - V^{(t)} \geq 0$ , respectivement. Les deuxième et quatrième colonnes décrivent les conditions équivalentes exprimées en fonction de  $n^{(t)} = n - l$ ,  $R(t; \gamma - \lambda/2)$ , et  $R(t; \lambda/2)$  correspondant à  $V^{(t)} - V^{(t+1)} < 0$  et  $V^{(t)} - V^{(t+1)} \geq 0$ , respectivement. L'une des conditions importantes que doit satisfaire la solution de l'algorithme de transfert  $l^*(\lambda)$  est que l'inégalité  $\pi^{(N-l)}(\lambda) < 1$  soit satisfaite. Il est facile de vérifier que  $\pi^{(N-l)}(\lambda) < 1 \Rightarrow A(t) > 0$ . En ce qui concerne la description de l'algorithme de transfert donnée au tableau A.1, cette condition signifie qu'il ne peut y avoir de solution que quand  $A(t) > 0$  aussi bien que  $B(t) \geq 0$  ou, de façon équivalente, quand  $n - l$  satisfait la condition (T.2).

Tous les états possibles définis par l'algorithme de transfert sont énumérés au tableau A.1. La correspondance entre les quantités des cases internes (calculables directement à partir des données auxiliaires et des paramètres estimés) et

précision. Étant donné le plan d'échantillonnage  $p^q(s, \lambda, l^q_*(\lambda, n^q))$ , la taille actuelle de l'échantillon  $n^q$  et la partition de la population  $\{U^{aq}_*(l^q_*)\}$ , on peut exprimer formellement la contrainte de précision imposée à  $t^q_R = t^q_a + t^q_{Rb}$  par la relation

$$CV^{min}_q \geq \frac{\hat{V}^{1/2}_q(t^q_{Rb}, \lambda, N - l^q_* - l^q_*)}{\hat{t}^q_R} \quad (4.1)$$

En résolvant l'inégalité pour  $n^{bq}_*$ , on obtient l'expression de la taille minimale que doit avoir l'échantillon dans le groupe échantillon  $U^{bq}_*(l^q_*)$  pour satisfaire la contrainte de précision, soit:

$$n^{bq}_* = n^q_* - l^q_*(n^q) = \frac{t^q_R CV^{min}_q + X(l^q_*, \hat{\gamma}) \varepsilon}{X(l^q_*, \lambda/2) X(l^q_*, \hat{\gamma} - \lambda/2) \varepsilon} \quad (4.2)$$

où  $X(l^q_*, \lambda/2) = \sum_{k=1}^{N-l^q_*-\lambda/2} x^{(k)}_{\hat{\gamma}-\lambda/2}$ ,  $X(l^q_*, \hat{\gamma} - \lambda/2) = \sum_{k=1}^{N-l^q_*-\hat{\gamma}-\lambda/2} x^{(k)}_{\hat{\gamma}-\lambda/2}$ , et où  $t^q_R$  peut être estimé à partir de données d'enquête antérieures correspondant à la période de référence des données auxiliaires. La nouvelle taille totale minimale de l'échantillon nécessaire pour satisfaire la contrainte de précision est alors donnée par

$$n^q_* = n^{aq}_* + n^{bq}_* = l^q_*(n^q) + n^{bq}_* \quad (4.3)$$

## 4.2 Méthode combinée de remaniement du plan d'échantillonnage

On notera maintenant que la solution de l'algorithme de transfert  $l^q_*$  dépend de la taille totale actuelle de l'échantillon, soit  $l^q_*(\lambda) = l^q_*(\lambda, n^q)$ . Une fois qu'on a déterminé la nouvelle taille minimale de l'échantillon  $n^q_*$ , la partition existante  $\{U^{aq}_*(l^q_*)\}$ , qui était optimale pour  $n^q_*$ , n'est plus optimale, car  $l^q_*(\lambda, n^q) \neq l^q_*(\lambda, n^q_*)$ . Par conséquent, si on pose  $n^{q+1}_* = n^q_*$ , il faut déterminer au moyen de l'algorithme de transfert une nouvelle partition de la population basée sur  $l^{q+1}_*(\lambda, n^{q+1}_*)$ , donnée par  $\{U^{a,q+1}_*(l^{q+1}_*), U^{b,q+1}_*(l^{q+1}_*)\}$ , pour optimiser la construction des sous-populations à tirage complet et échantillon. Puis, l'application de (4.2) à  $U^{b,q+1}_*(l^{q+1}_*)$  donne la nouvelle taille minimale d'échantillon  $n^{q+1}_* = l^{q+1}_*(n^{q+1}_*) + n^{b,q+1}_*$  nécessaire pour satisfaire la contrainte de précision  $CV^{min}$  imposée. En procédant de cette façon, la méthode combinée produit une série de partitions de population, de tailles d'échantillons et de répartitions d'échantillons représentée par

$$(l^q_*(\lambda, n^q), (n^{aq}_* = l^q_*(\lambda, n^q), n^{bq}_* = n^q - l^q_*)) \quad (4.4)$$

où  $n^{q+1}_* = n^q_* + n^{bq}_*$  et où la valeur initiale est  $n_0$  (taille actuelle de l'échantillon d'enquête). On répète la méthode combinée jusqu'à ce qu'on ne puisse plus réduire la taille minimale de l'échantillon. On arrive ainsi à la règle d'arrêt

$$q^* = \min\{q : n^{q+1}_* - n^q_* \geq 0\} \quad (4.5)$$

## 5. APPLICATION

La démonstration de l'optimalité de la méthode combinée, qui peut être faite au moyen du théorème 2, n'est pas exposée ici pour des raisons d'espace (voir Pandher 1995). Le résultat principal est que la méthode combinée tend vers une solution optimale globale selon le cheminement défini par (4.4) et finit par aboutir à un stade où il n'est plus possible de réduire la taille de l'échantillon (en reconstruisant  $U^a_*$  et  $U^b_*$ ) étant donné la contrainte de précision imposée.

Appliquons maintenant la méthode combinée d'établissement du plan d'échantillonnage décrite plus haut au remaniement de l'Enquête sur les finances des administrations locales (EFAL) effectuée en Ontario. Dans cette application, les données d'enquête  $y$  correspondent aux dépenses réelles déclarées par les administrations locales (unités) échantillonnées en Ontario en 1989. Les estimations réelles sont préparées 30 mois après la fin de l'année d'enquête à partir des états financiers présentés par les administrations locales au ministère provincial des Affaires municipales. La variable auxiliaire  $x$  correspond aux dénombrements de la population d'administrations locales du recensement le plus rapproché dans le temps (1991). En Ontario, la population d'administrations locales compte en tout 793 unités dont on tire, à l'heure actuelle, un échantillon de 108 unités.

Les résultats de l'application de la méthode combinée aux données de l'EFAL effectuée en Ontario sont présentés dans le tableau 1. La contrainte de précision  $VC^{min}$  a été fixée à 2% pour l'estimateur de régression du total  $t^q_R = t^q_a + t^q_{Rb}$ . En appliquant les méthodes de Pandher (1995), on a déterminé que, pour l'Ontario, la meilleure valeur du paramètre d'hétéroscédasticité  $\gamma$  est  $\hat{\gamma} = 2$ ; on a aussi calculé la valeur de la constante de proportionnalité correspondante, soit  $\hat{c} = .0825$ . Enfin, on a utilisé le plan d'échantillonnage presque optimal défini par  $\lambda = \hat{\gamma}$  ( $p(s; \hat{\gamma})$ ).

Tableau 1  
Application de la méthode combinée aux données de l'EFAL (Ontario, 1989)

Itération (q)		$n^q$	$l^q_*(\lambda, n^q)$	$n^{aq}_*$	$n^{bq}_*$	$n^q_*$
0	108	39	12	12	38	50
1	57	16	16	16	34	50
2	50	18	39	39	18	57

Pour l'Ontario, l'application de la méthode combinée s'est terminée à l'itération  $q^* = 2$ . La partition globalement optimale de la population en un groupe à tirage complet et en un groupe échantillon correspond à  $N^a_* = 16$  et  $N^b_* = 777$ . La nouvelle taille minimale de l'échantillon total est  $n^* = 50$ , avec les allocations  $n^a_* = 16$  et  $n^b_* = 34$ . Donc, on obtient une réduction de la taille de l'échantillon total de  $n_0 - n^*_\lambda = 108 - 50 = 58$  pour l'estimateur de régression  $t^q_R = t^q_a + t^q_{Rb}$  ayant un coefficient de variation de 2%.



$(n^{(i-1)})^b = n^{(i)}$  à l'effet inverse, autrement dit, augmentant  $V^{(i)}(t_{R^b}; \cdot)$ . À un point donné du processus existe une valeur critique,  $t^*$ ,  $0 \leq t^* < n$  qui donne la répartition optimale  $\{U_a^*(t^*), U_b^*(t^*)\}$ . Qui plus est, dans l'exposé du Théorème 3 de Pandher (1995), on montre qu'il existe une solution de l'algorithme de transfert et que le système reste stable (optimal) quant il a atteint  $t^*$  aussi longtemps que les conditions  $(x^{(N-i)}_{N-i} - x^{(N-i-1)}_{N-i-1}) \geq 0$  et  $(x^{(N-i)}_{N-i} - x^{(N-i-1)}_{N-i-1}) \geq 0$ ,  $0 \leq t < n$  sont satisfaites. La stabilité confirme que la solution est optimale, puisque les conditions menant à la solution ne varient pas dans l'intervalle  $t^* \leq t < n$ . Ces deux propriétés se définissent plus précisément de la façon suivante:

existence:  $\exists t^*, 0 \leq t^* < n$ , de sorte que  $V^{(i+1)} - V^{(i)} \geq 0$  et  $\pi^{(N-i-1)} < 1$ ;

stabilité: Si  $V^{(i+1)} - V^{(i)} \geq 0$ , alors  $V^{(i+1)} - V^{(i)} \geq 0$  et  $\pi^{(N-i-1)} < 1$  pour  $0 \leq t^* < l < n$ .

Un exemple de l'application de l'algorithme de transfert à la population de l'Enquête sur les finances des administrations locales en Ontario (avec  $N = 793$ ,  $n = 108$ ,  $\gamma = 2$ , et  $\lambda = 1$ ) est donné à la figure 1. Les courbes sont tracées pour  $l > 8$  car dans l'intervalle  $0 < l \leq 8$ , la première condition de (3.3), à savoir  $[\pi^{(N-l)}(\lambda) < 1]$ , n'est pas satisfaite. La valeur minimale de  $V^{(i)}(t_{R^b})$  est obtenue pour  $t^* = 57$  où  $\Delta(t^*) = V^{(i+1)} - V^{(i)} \geq 0$ .

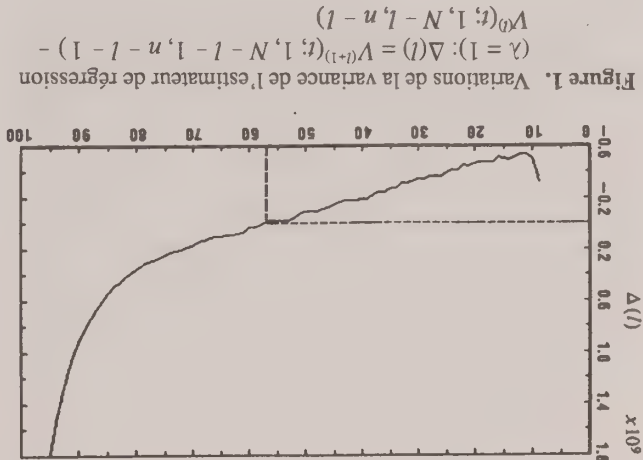


Figure 1. Variations de la variance de l'estimateur de régression

La solution  $t^*(\lambda)$  de l'algorithme de transfert énoncée au moyen de l'équation (3.3) en fonction de  $V^{(i)} - V^{(i-1)}$  et  $\pi^{(N-i)}(\lambda)$  peut aussi être exprimée de façon équivalente par

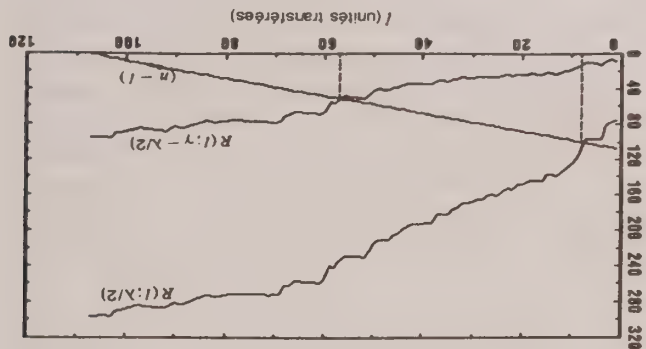
#### transfert

### Théorème 2. Solution équivalente de l'algorithme de

Le Théorème 2 décrit dans l'article complet est un résultat important qui permet d'exprimer la solution de l'algorithme de transfert de façon équivalente au moyen de quantités plus simples, fondées sur les données auxiliaires. Une brève description du développement de ce théorème figure en annexe.

### 4. DÉTERMINATION DE LA TAILLE DE L'ÉCHANTILLON ET MÉTHODE ITÉRATIVE COMBINÉE

Figure 2. Utilisation de  $R(l; \gamma - \lambda/2)$ ,  $R(l; \lambda/2)$ , et  $(n - l)$  pour construire les groupes à tirage complet et échantillon optimaux (Ontario)



obtient la même solution  $t^* = 57$  que précédemment.  $R(l; \gamma - \lambda/2)$  (la courbe inférieure du graphique) et  $n - l$ . On la solution est déterminée par le comportement des fonctions figure 2 (données pour l'Ontario). Dans ce cas,  $0 \leq \lambda < \gamma$ , et optimale de la population est illustrée graphiquement à la application de ce théorème pour déterminer la répartition  $\sum_{k=1}^{N-l} x^{(k)}_{N-l} / x^{(k)}_{N-l}$  définissent les valeurs critiques.

où  $R(l; \gamma - \lambda/2) = \sum_{k=1}^{N-l} x^{(k)}_{N-l} / x^{(k)}_{N-l}$  et  $R(l; \lambda/2) =$

$$t^*(\lambda) = \begin{cases} \min \{l : n - l \leq R(l; \lambda/2), 0 \leq l < n\}, & \gamma < \lambda \leq 2\gamma \\ \min \{l : n - l \leq R(l; \gamma/2), 0 \leq l < n\}, & \lambda = \gamma \\ \min \{l : n - l \leq R(l; \gamma - \lambda/2), 0 \leq l < n\}, & 0 \leq \lambda < \gamma \end{cases}$$

Étant donné un plan d'échantillonnage  $p(\lambda)$ ,  $0 \leq \lambda \leq 2\gamma$ , avec une taille d'échantillon  $n$ , l'algorithme de transfert produit une construction optimale des sous-populations à tirage complet et échantillon,  $U_a^*(t^*)$  et  $U_b^*(t^*)$ , respectivement. Puis, on définit une expression qui permet de déterminer la taille minimale de l'échantillon en satisfaisant les contraintes de précision imposées, lesquelles sont exprimées en fonction du coefficient de variation  $CV^{min}$ . Ensuite, on intègre l'étape de détermination de l'échantillon à l'algorithme de transfert pour élaborer une méthode combinée qui permet au concepteur de l'enquête de déterminer la taille globalement minimale de l'échantillon et la partition globalement optimale de la population.

#### 4.1 Expression de la nouvelle taille de l'échantillon

Représentons par  $q$  le cycle d'itération de la méthode combinée et par  $n_q^* = n_{aq}^* + n_{bq}^*$ , la taille totale minimale de l'échantillon nécessaire pour satisfaire la contrainte de

où  $t_a^* = \sum_{k \in U_p} x_k$ , puisque toutes les unités du groupe à tirage complet sont échantillonnées et que  $t_{Rp}$  représente l'estimateur de régression généralisé correspondant au modèle pertinent. La variance attendue de  $\hat{t}_{Rp}$  (définie comme étant la variance liée tant au plan d'échantillonnage qu'au modèle, représentées par  $p$  et  $\xi$ , respectivement) peut être exprimée par la relation:

$$V(\hat{t}_{Rp}) \equiv E_{\xi} V_p(\hat{t}_{Rp}) = \sum_{k \in U_p} \left( \frac{1}{1} - 1 \right) \sigma_k^2 \quad (2.3)$$

En outre, si  $\sigma_k^2$  dépend de la mesure auxiliaire  $x_k$  conformément à l'équation (2.4), alors l'optimalité du plan d'échantillonnage (Godambe et Joshi 1965) implique que les probabilités d'inclusion dans l'échantillon optimal sont  $\pi_k^*(\gamma) \propto x_k^{\gamma/2}$ ,  $k \in U_p$ . Donc, pour la sous-population échantillonnée, le plan d'échantillonnage  $p_b^*(s_p, \gamma)$ , qui définit les probabilités d'inclusion de premier ordre  $\pi_k^*(\gamma) = n(x_k^{\gamma/2} / \sum_{j \in U_p} x_j^{\gamma/2})$ ,  $k \in U_p$ , minimise la variance attendue  $V(\hat{t}_{Rp})$ .

Dans le cas du cadre théorique assisté par modèle utilisé dans le présent article, on suppose que la mesure auxiliaire  $x_k$  est scalaire. Comme l'a souligné un examinateur, on pourrait traiter le cas plus général où  $x_k$  est un vecteur en ajustant la relation paramétrique appropriée  $\sigma_k^2 = f(x_{k1}, \dots, x_{kg})$  et en utilisant la valeur estimée  $\hat{\theta}_k$  au lieu de  $x_k$  pour définir les probabilités d'inclusion. La méthode proposée pour traiter la variable multidimensionnelle  $x_k$  qui paraît intuitivement valable, est mentionnée ici par souci de complétude, mais doit être étudiée plus en profondeur.

Trois méthodes d'estimation du paramètre d'hétéroscasticité  $\gamma$  à partir des données d'une enquête antérieure, à savoir la «Méthode des moindres carrés», la «Méthode du maximum de vraisemblance» et la «Méthode graphique» sont décrites à l'annexe A dans Pandher (1995).

### 3. ALGORITHME DE TRANSFERT

Dans la présente section, on propose un mécanisme itératif nommé «algorithme de transfert» pour délimiter de façon optimale les sous-populations à tirage complet et échantillon dans le cadre du plan d'échantillonnage  $p(s, \lambda)$ . Le critère de délimitation consiste à trouver une partition de la population qui réduit au minimum l'estimation de la variance attendue de  $\hat{t}_{Rp}$ . On se sert d'une solution équivalente donnée par Pandher (1995) pour obtenir une autre méthode, plus simple, de résolution de l'algorithme fondée entièrement sur des quantités définies d'après les données auxiliaires de la population.

Le mécanisme proposé pour construire les sous-populations à tirage complet et échantillon,  $U_a$  et  $U_b$ , respectivement, se fonde sur l'idée qui suit. Au départ, on place toutes les unités de la population dans le groupe échantillon qu'on désigne par  $U_{(0)}^b$  (l'indice supérieur  $l$  représente le cycle d'itération). Par conséquent, le groupe à tirage complet est un ensemble vide  $U_{(0)}^a = \{\emptyset\}$ . La population et la répartition de la taille de l'échantillon résultantes pour  $l = 0$  est

$$V_{(0)}^b(\hat{t}_{Rp}; \cdot) = \sum_{k \in U_{(0)}^b} \left( \frac{1}{1} - 1 \right) \pi_k^b(\lambda) = 0, \quad N_{(0)}^b = 0, \quad N_{(0)}^a = N, \quad \text{et } n_0^b = n_0, \quad \text{où } n_0 \text{ est la taille actuelle de l'échantillon.} \quad (3.1)$$

$V_{(0)}^b(\hat{t}_{Rp}; \cdot)$  suivant:

Dans l'algorithme itératif, on prend au départ toutes les unités de la population placées dans  $U_{(0)}^b$ . Puis, pour chaque itération  $l$ ,  $0 \leq l < n$ , on transfère l'unité  $x_{(N-l-1)}^b$  à  $U_{(l)}^b$  et on calcule la différence

$$\Delta(l) = V_{(l+1)}^b(\hat{t}_{Rp}; \lambda, N - l - 1, n - l - 1) - V_{(l)}^b(\hat{t}_{Rp}; \lambda, N - l, n - l) \quad (3.2)$$

Les valeurs négatives de  $\Delta(l)$  signifient que le transfert de l'unité correspondant à la valeur ordonnée  $x_{(N-l-1)}^b$  fait diminuer la variance. En outre, un tel transfert continue de faire diminuer la variance de  $\hat{t}_{Rp}$ , aussi longtemps que  $\Delta(l) < 0$ . En général, pour toute itération  $l$ , la relation entre la population et les répartitions de la taille de l'échantillon est représentée par les relations  $N_{(l)}^b = N - l$ ,  $n_{(l)}^b = n - l$ , et  $N_{(l)}^a = n_{(l)}^a = l$ . Ces relations sont satisfaites parce que les tailles de la population totale et de l'échantillon doivent demeurer constantes ( $N = N_{(l)}^a + N_{(l)}^b$  et  $n = n_{(l)}^a + n_{(l)}^b$ ) pour toutes les itérations.

La solution est également soumise à la contrainte  $\pi_k^b(\lambda) < 1$ ,  $k \in U_{(l)}^b$ . Supposons que  $l^* (\lambda)$ ,  $0 \leq l^* < n$ , représente la solution de l'algorithme de transfert. Compte tenu de la discussion qui précède, la solution de l'algorithme de transfert pour le plan d'échantillonnage  $p(s, \lambda)$  peut s'écrire

$$l^*(\lambda) = \min \{ l : \pi_{(N-l)}^b(\lambda) > 1 \} \quad \text{et} \quad \Delta(l) = [V_{(l+1)}^b(\hat{t}_{Rp}; \lambda) - V_{(l)}^b(\hat{t}_{Rp}; \lambda)] \geq 0, \quad 0 \leq l < n. \quad (3.3)$$

L'allocation optimale de la population au groupe à tirage complet  $U_a^*(l^*)$  correspond alors aux unités de population qui coïncident avec les unités d'ordre  $l^*$  transférées au vecteur auxiliaire à tirage complet  $X_a^* = (x_{(N-l^*-1)}^b, \dots, x_{(N-l^*-1)}^b)$ ; de façon similaire, le groupe échantillon  $U_b^*(l^*)$  comprend les unités correspondant à  $X_b^* = (x_{(1)}^b, x_{(2)}^b, \dots, x_{(N-l^*-1)}^b)$ . Le transfert d'une unité de  $U_{(l)}^b$  à  $U_{(l+1)}^b$  a deux effets opposés sur la variance  $V_{(l+1)}^b(\hat{t}_{Rp}; \cdot)$ . La réduction de la taille de la population ( $N_{(l+1)}^b = N_{(l)}^b - 1$ ) fait diminuer la variance, tandis que la réduction équivalente de la taille de l'échantillon



## 2) Choix d'une méthode efficace de sélection de l'échantillon

Posons que  $p(s; \lambda) = (p_a(s_a), p_b(s_b; \lambda))$  représente le plan d'échantillonnage complet, où le paramètre  $\lambda$  du plan d'échantillonnage détermine le type d'échantillonnage exécuté pour le groupe échantillon  $U_b$ . On peut exprimer les probabilités d'inclusion dans l'échantillon étant donné  $p_a(s_a, \lambda)$  par l'expression  $\pi_k(\lambda) = n_b(x_k) / \sum_{j \in U_b} x_j$ ,  $k \in U_b$ . Il convient de souligner que le paramètre  $\lambda$  définit une classe générale de plans d'échantillonnage dont l'échantillonnage aléatoire simple ou EAS ( $\lambda = 0$ ) et l'échantillonnage avec probabilité proportionnelle à la taille ou PPT ( $\lambda = 2$ ) sont des cas particuliers. Les résultats sur l'optimalité du plan d'échantillonnage (Godambe et Joshi 1965) permettent de déterminer la valeur la plus optimale pour le paramètre de plan d'échantillonnage  $\lambda$ .

### 3) Détermination de la taille minimale de l'échantillon

Le troisième élément de la méthodologie globale consiste à déterminer la taille minimale de l'échantillon nécessaire pour satisfaire les contraintes de précision imposées pour l'estimateur.

La méthode combinée que l'on propose intègre ces éléments pour permettre de déterminer une nouvelle taille globalement minimale d'échantillon et une nouvelle partition globalement optimale de la population pour une gamme souple de méthodes d'échantillonnage (p. ex., EAS, PPT, PPT généralisé). Pour commencer, on propose l'algorithme de transfert qui permet de définir la répartition optimale de la population entre le groupe «à tirage complet» et le groupe «échantillon» de façon à minimiser la variance de l'estimateur de régression généralisée du total. Les propriétés mathématiques souhaitables de cet algorithme, dont l'existence d'une solution et l'optimalité de cette dernière, ainsi qu'une solution équivalente, ont été établies par Pandher (1995). La solution équivalente peut être exprimée au moyen de quantités simples, calculables directement à partir des données auxiliaires sur la population.

Puis, on synthétise, selon une méthode itérative, l'algorithme de transfert et l'étape de détermination de la taille de l'échantillon pour trouver la taille minimale de l'échantillon nécessaire pour satisfaire les contraintes de précision imposées. La méthode combinée produit une série de tailles d'échantillon et de partitions de population qui tendent vers une solution globalement optimale correspondant au point où une réduction ultérieure de la taille de l'échantillon devient insupportable compte tenu des contraintes de précision imposées. On illustre l'application de la méthode au moyen de données provinciales tirées de l'Enquête sur les finances des administrations locales en Ontario.

Lavallée et Hidiroglou (1988), Hidiroglou et Srinath (1993) (subéquemment désignés par L&H et H&S, respectivement) et Glasser (1962) ont proposé d'autres méthodes de construction des groupes à tirage complet et échantillon dans le contexte d'un plan d'échantillonnage aléatoire simple stratifié. La méthode proposée ici se distingue des leurs à trois égards. Premièrement, la partition de la population est définie pour toute une gamme de méthodes d'échantillonnage

## 2. CADRE THÉORIQUE D'ENQUÊTE

On adopte le cadre théorique d'enquête assisté par modèle pour la population à distribution asymétrique dont les variables auxiliaires et les variables étudiées sont représentées par  $C_U = \{(x_1, y_1), \dots, (x_N, y_N)\}$ . Ce cadre comprend des estimateurs de régression généralisés de la population totale appuyés sur des modèles de régression exploitant la corrélation entre les variables étudiées  $y$  (Sæmdal 1992, p. 255) et les variables auxiliaires  $x$ . Les estimateurs varient selon les hypothèses émises quant aux composantes déterministes et stochastiques du modèle sous-jacent. Par exemple, un modèle hétéroscédastique proportionnel

$$y_k = \beta x_k + \epsilon_k, \quad (2.1)$$

dont l'erreur est  $\epsilon_k \sim (0, \sigma_k^2)$  et dont la structure de la variance est donnée par  $\sigma_k^2 = c x_k^\gamma$  ( $\gamma$  étant le paramètre d'hétéroscédasticité) produit l'estimateur de régression généralisé suivant:

$$\hat{t}_{Rb} = \sum_{k=1}^{U_b} x_k \hat{b} + \sum_{k=1}^{s_b} \frac{\pi_k}{(y_k - x_k \hat{b})} \quad (2.2)$$

où  $\hat{b} = (\sum_{k=1}^s y_k / \pi_k) / (\sum_{k=1}^s x_k / \pi_k)$  est l'estimation à pondération probabiliste basée sur l'échantillon du paramètre de régression de la population  $B$ .

Étant donné ce cadre théorique d'estimation, l'estimation du total des deux groupes  $t = t_a + t_b$  correspond à  $\hat{t} = \hat{t}_a + \hat{t}_{Rb}$



# Remaniement optimal du plan d'échantillonnage pour une population à distribution asymétrique au moyen d'un estimateur de régression généralisé avec applications

GURUPDESH S. PANDHER<sup>1</sup>

## RÉSUMÉ

On présente une méthode combinée, élaborée dans le cadre d'un remaniement d'enquête, permettant de déterminer la taille minimale de l'échantillon requise pour appliquer l'estimateur de régression généralisé à une population à distribution asymétrique (p. ex., entreprises, institutions, entreprises agricoles). Pour être efficace, la stratégie de remaniement du plan d'échantillonnage doit comprendre trois éléments, à savoir: i) la partition efficace de la population en un groupe «à tirage complet» et un groupe «échantillon», ii) la définition d'une méthode efficace de sélection de l'échantillon et iii) la détermination de la taille minimale de l'échantillon requise pour satisfaire la ou les contraintes de précision imposées. On conçoit un mécanisme, baptisé «algorithme de transfert», pour résoudre la première question (Pandher 1995) et on l'intègre aux deux autres éléments pour aboutir à une méthode itérative combinée qui converge vers une taille globalement minimale d'échantillon et une partition de la population qui satisfait les contraintes de précision imposées. Parallèlement, on obtient une solution équivalente de l'algorithme de transfert qu'on peut exprimer au moyen de quantités simples, calculables directement à partir des données auxiliaires sur la population. Enfin, on présente les résultats de l'application de la méthode proposée de remaniement de l'échantillon à l'enquête sur les finances des administrations locales en Ontario. On obtient une réduction de 52% de la taille globale de l'échantillon quand on fixe à 2% la valeur minimale du coefficient de variation pour l'estimateur de régression généralisé du total.

**MOTS CLÉS:** Taille minimale de l'échantillon; sélection optimale de l'échantillon; contrainte de précision; groupe d'échantillon; groupe à tirage complet.

## 1. INTRODUCTION

Dans le cas de nombreuses enquêtes, on dispose de renseignements supplémentaires sur toutes les unités de la population avant l'exécution. On se sert souvent de ces données auxiliaires pour concevoir un plan d'échantillonnage et une stratégie d'estimation plus efficaces. Dans le contexte du remaniement d'une enquête, la stratégie optimale est celle qui promet d'offrir la plus forte réduction des coûts d'enquête grâce à l'utilisation de l'échantillon ayant la taille minimale requise pour satisfaire les contraintes de précision imposées. Dans le cas d'enquêtes répétées sur des populations à distribution asymétrique, on aboutit à un plan d'échantillonnage et à une stratégie d'estimation efficaces en exploitant a) la corrélation entre les données auxiliaires sur la taille  $x$  (p. ex., population de la municipalité, effectif de l'entreprise, super-ficte de l'entreprise agricole) et les variables étudiées  $y$  (p. ex., dépenses de la municipalité, valeur des expéditions, rendement agricole) et b) la covariance des variables étudiées et des données auxiliaires sur la taille.

Dans le présent article, on décrit l'élaboration d'une méthode générale de remaniement du plan d'échantillonnage pour les populations à distribution asymétrique ayant pour objectif ultime de réduire au maximum la taille de l'échantillon existant tout en veillant à satisfaire les contraintes de précision établies pour l'estimateur de régression généralisé du total. Ces travaux ont été entrepris à l'occasion du remaniement de l'enquête sur les finances des administrations

Puisque la variabilité de la réponse à l'enquête  $y_k$  a tendance à augmenter parallèlement à la taille de l'unité  $x_k$ , il est courant, dans le cas des populations à distribution asymétrique, d'échantillonner avec certitude les unités correspondantes aux valeurs de  $x$  les plus grandes afin d'améliorer l'efficacité des estimateurs de population. On effectue la partition de la population en deux groupes qui ne se chevauchent pas, à savoir le groupe «à tirage complet»  $U_a = \{1, \dots, N_a\}$  et le groupe «échantillon»  $U_b = \{1, \dots, N_b\}$ , grâce à un nouveau mécanisme baptisé «algorithme de transfert».

### 1) Création des groupes «à tirage complet» et «échantillon»

En vue de définir un nouveau plan d'échantillonnage efficace, la méthodologie globale fournit et intègre les solutions de trois problèmes:

(p. ex., enquêtes sur les entreprises agricoles, sur les entreprises et sur les institutions).

En vue de définir un nouveau plan d'échantillonnage efficace, la méthodologie globale fournit et intègre les solutions de trois problèmes:

locales effectuée par la Division des institutions publiques de Statistique Canada. Les renseignements financiers (p. ex., revenus, dépenses, dettes, etc.) fournis par les administrations locales servent à calculer et à publier des statistiques financières aux échelons provincial et national. Bien que les travaux présentés ici aient été effectués dans le cadre d'une application concrète, la méthode d'établissement du plan d'échantillonnage proposée peut être généralisée à toutes les enquêtes sur des populations à distribution asymétrique (p. ex., enquêtes sur les entreprises agricoles, sur les entreprises et sur les institutions).

<sup>1</sup> Gurupdes S. Pandher, Section du développement d'enquêtes et d'analyse, Division des méthodes d'enquêtes des ménages, Direction de la méthodologie, Statistique Canada, 16<sup>e</sup> étage, immeuble R.H. Coats, Ottawa, Ontario, Canada, K1A 0T6.



## BIBLIOGRAPHIE

- BRICK, J., BURKE, J., et WEST, J. (1992). Telephone Under-coverage Bias of 14- to 21-year-olds and 3- to 5-year-olds. National Household Education Survey Technical Report No. 2, Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, NCEES 92-101.
- BRICK, J., KEETER, S., WAKSBERG, J., et BELL, B. (1996). Adjusting for Coverage Bias Using Telephone Service Interruption Data. National Household Education Survey Technical Report, Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, NCEES 96-336.
- COCHRAN, W. (1977). *Sampling Techniques*. New York: John Wiley and Sons, 12-15.
- CONVERSE, J., et PRESSER, S. (1986). *Survey Questions, Handcrafting the Standardized Questionnaire*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-063. Beverly Hills: Sage Publishers.
- FEDERAL COMMUNICATIONS COMMISSION (1988). Monitoring Report: CC Docket No. 87-339. Préparé par le personnel du comité conjoint Fédéral-Etats dans CC Docket No. 80-286, Washington DC.
- KALTON, G., et KASPRZYK, D. (1986). Le traitement des données d'enquête manquant. *Techniques d'enquête*, 12, 1-17.
- KEETER, S. (1995). Estimating noncoverage bias from a phone survey. *Public Opinion Quarterly*, 59, 196-217.
- KISH, L. (1992). Weighing for unequal  $P_i$ . *Journal of Official Statistics*, 8, 183-200.
- MASSÉ, J., et BOTMAN, S. (1988). Weighing adjustments for random digit dialed surveys, Chapitre 9 dans *Telephone Survey Methodology*. Eds. Groves, Biemer, Lyberg, Massey, Wakseberg. New York: John Wiley and Sons, 143-160.
- POLITZ, A., et SIMMONS, W. (1949). An attempt to get the 'not at homes' into the sample without callbacks. *Journal of the American Statistical Association*, 44, 9-31.
- THORNBERRY, O., et MASSÉ, J. (1988). Trends in United States telephone coverage across time and subgroups, Chapitre 3 dans *Telephone Survey Methodology*. Eds. Groves, Biemer, Lyberg, Massey, Nicholls, et Wakseberg. New York: John Wiley and Sons, 25-50.

Les autres méthodes de pondération ont donné des résultats différents pour le ratio quadratique moyen. Celles articulées sur une interruption du service téléphonique d'une semaine fonctionnent mieux que celles reposant sur une interruption supérieure à un mois. L'ajustement du biais au moyen du degré de scolarité selon la race est à peu près équivalent à celui obtenu avec le statut de propriété selon la race. On devrait tenir compte de la taille de l'échantillon lorsqu'on envisage un ajustement d'après l'interruption du service téléphonique. En effet, le ratio du biais augmente avec la taille de l'échantillon, car cette dernière ne modifie pas le biais mais bien l'erreur d'échantillonnage de l'estimation (dénominateur du quotient). Par conséquent, l'ajustement devrait s'avérer plus utile dans les enquêtes reposant sur un échantillon important, où le ratio du biais devrait être passablement important.

Bien que les résultats de l'étude tendent à démontrer l'utilité de l'ajustement pour bon nombre d'estimations issues des sondages téléphoniques, il n'en reste pas moins qu'on doit les confirmer avant de les mettre en application. Comme nous l'avons indiqué précédemment, les estimations de l'erreur quadratique moyenne partent de l'hypothèse que l'ajustement de l'estimation élimine le biais que peut renfermer celle-ci. Or, cette hypothèse n'a pu être vérifiée, faute de données de référence. Le modèle expérimental améliore considérablement l'estimation ajustée, en ce sens qu'il baisse les limites de l'erreur quadratique moyenne. Par conséquent, les résultats de l'étude devraient être considérés comme un signe que l'ajustement au moyen des interruptions du service téléphonique est réalisable. Cette méthode nécessiterait toutefois une étude et une évaluation plus poussées.

Des questions sur l'interruption du service téléphonique ont récemment été ajoutées à l'interview de la National Health Interview Survey entreprise par le Census Bureau au nom du National Center for Health Statistics. Les résultats de cette enquête devraient s'avérer très utiles pour évaluer la méthode proposée, car l'enquête couvre les ménages sans téléphone au moyen d'une interview directe, ce qui élimine la nécessité de recourir à l'hypothèse primordiale du modèle présentée ici.

## REMERCIEMENTS

Les auteurs tiennent à remercier les examinateurs et le rédacteur en chef pour leurs commentaires qui ont contribué à améliorer sensiblement la méthodologie et la présentation du document.



Tableau 5  
Sommaire de la distribution du ratio quadratique moyen pour certaines caractéristiques des populations, préparation à l'école et sécurité et discipline à l'école

Méthode d'ajustement	Sécurité et discipline à l'école				Préparation à l'école			
	A1	A2	B1	B2	Moyenne	Médiane	Minimum	Maximum
Méthode d'ajustement					89.8	96.0	27.0	120.0
					101.0	108.0	30.3	135.0
					86.8	92.8	26.1	116.0
					94.2	100.8	28.3	126.0
Méthode d'ajustement					93.3	100.8	26.4	112.0
					104.9	113.4	29.7	126.0
					92.2	99.9	26.2	111.0
					103.9	112.5	29.5	125.0

Source: U.S. Department of Education, National Center for Education Statistics, National Household Education Survey, printemps 1993.

et une moyenne supérieures à 100. Bref, le ratio est déplacé vers le haut comparativement à celui des méthodes A1 et B1. Son utilisation n'est donc pas recommandée.

## 5. CONCLUSIONS

Si le pourcentage de la population visée, qui se retrouve dans les ménages sans téléphone, est relativement élevé et si les caractéristiques de ces personnes diffèrent de celles des personnes vivant dans un ménage avec téléphone, il se pourrait que les estimations soient faussées par un biais de couverture important. Une façon de résoudre le problème sans recourir à d'autres méthodes de collecte de données consiste à ajuster les poids en vue d'atténuer le biais. Dans la présente étude, nous avons majoré les poids des personnes faisant partie d'un ménage où l'on avait signalé une interruption du service téléphonique pour tenir compte de la population sans téléphone.

La réduction estimée du biais calculée grâce au modèle expérimental révèle que l'ajustement de la couverture pour la population PE améliore sensiblement une partie des estimations sans nuire de façon appréciable aux statistiques. La réduction estimée du biais pour la population SDE, en revanche, n'est pas aussi importante. Les ajustements atténuent le biais pour les deux populations, mais ils augmentent aussi la variabilité des estimations. La distribution du ratio quadratique moyen révèle qu'on pourrait améliorer environ la moitié des estimations si on procédait à un ajustement d'après l'interruption du service téléphonique. Par ailleurs, la variance n'augmente pas beaucoup, y compris pour les estimations qui perdent de leur précision consécutivement à l'utilisation de poids différents. En d'autres termes, l'ajustement ne cause pas une grande perte même quand il ne diminue pas le biais de couverture. Ces constatations donnent à penser qu'on devrait sérieusement envisager de recourir à un tel ajustement.

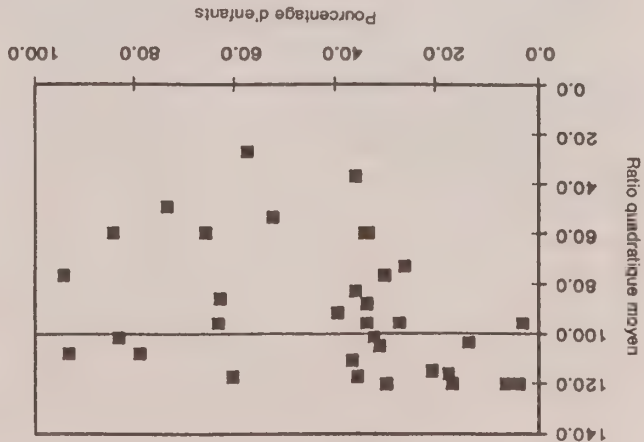


Figure 4. Ratio quadratique moyen estimé de certaines caractéristiques de la population «Préparation à l'école» (méthode A1)

Source: U.S. Department of Education, National Center for Education Statistics, National Household Education Survey, printemps 1993.

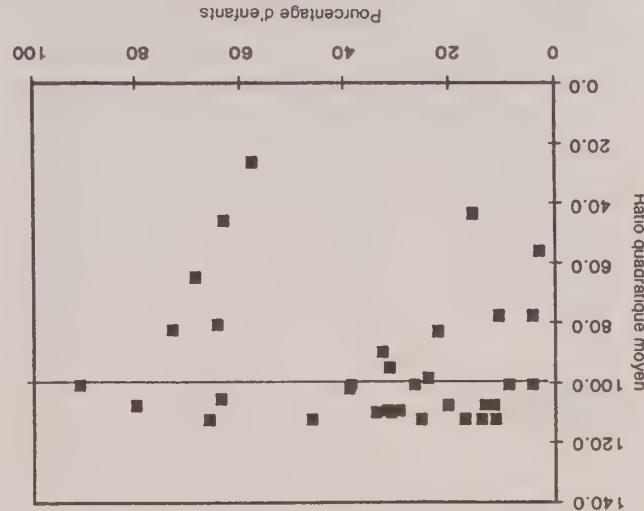


Figure 5. Ratio quadratique moyen estimé de certaines caractéristiques de la population «Sécurité et discipline à l'école» (méthode A1)

Source: U.S. Department of Education, National Center for Education Statistics, National Household Education Survey, printemps 1993.

lorsque cette dernière n'est pas biaisée, ne présente donc qu'un faible inconvénient comparativement aux avantages, tout à fait impressionnants. Le ratio quadratique moyen est distribué de façon très semblable pour les méthodes A1 et B1. Le choix entre l'une ou l'autre méthode pourrait donc reposer sur des considérations non statistiques, l'existence de données et les autres types d'ajustement qu'exige l'enquête, par exemple. Le ratio quadratique moyen indique que les poids ajustés réduisent l'erreur quadratique moyenne pour environ la moitié des estimations qu'on estime être inférieures à celles obtenues avec les poids types. La distribution du ratio quadratique moyen des méthodes A2 et B2 (qui supposent une interruption du service téléphonique d'un mois ou davantage) a une médiane

On compare ensuite l'erreur quadratique moyenne des estimations ajustées à la variabilité des estimations types de la NHES:93. L'augmentation de la variance qui résulte de l'ajustement des poids par les données sur l'interruption du service téléphonique est exprimée sous la forme d'un *FIV* au tableau 3. En multipliant la variance estimée des estimations types par le bon coefficient d'ajustement, on obtient la variance approximative des estimations ajustées (qu'on suppose non biaisées). Il suffit ensuite de comparer ces dernières à l'erreur quadratique moyenne des estimations type.

Pour rendre la comparaison des méthodes de pondération plus facile, le ratio de la variance des estimations ajustées et de l'erreur quadratique moyenne de l'estimation type est présenté sous forme de tableau (lire Brick et coll. 1996). Ce ratio, baptisé ratio quadratique moyen, peut être exprimé comme suit:

$$(6) \quad \text{rqm}^a(p) = \frac{\text{eqm}^a(p)}{100 \times \text{relatif } FIV^a \times \text{var}(p_s)}$$

Remarquons qu'on dérive l'erreur quadratique moyenne du biais estimé uniquement de la méthode A1, tandis que le ratio quadratique moyen s'applique aux quatre méthodes. Ainsi que nous l'avons indiqué précédemment, pareille simplification n'a pas un effet appréciable sur le ratio quadratique moyen, car le biais estimé est à peu près le même d'une méthode à l'autre.

Le ratio quadratique moyen incorpore l'effet du biais (dans l'estimation de l'erreur quadratique moyenne) et de la variance (dans le *FIV*). Quand le ratio quadratique moyen est égal à 100, la variance de l'estimation ajustée est identique à l'erreur quadratique moyenne de l'estimation type biaisée. Un ratio inférieur à 100 indique que la réduction du biais réalisée grâce à l'ajustement dépasse la hausse de la variance qui s'y associe. Un ratio quadratique moyen (rqm) supérieur à 100 signifie que la hausse de la variance qui accompagne l'ajustement est plus importante que la réduction du biais obtenue.

Les figures 4 et 5 présentent de façon graphique le rqm des deux enquêtes selon la méthode A1. De son côté le tableau 5 donne les statistiques sommaires relatives au rqm pour les quatre méthodes d'ajustement. La distribution du ratio quadratique moyen se ressemble pour les deux groupes, quoique le ratio quadratique moyen soit légèrement plus faible pour la population PE. Les médianes des méthodes A1 et B1 (qui supposent une interruption du service téléphonique d'au moins une semaine) se trouvent presque au point d'équilibre, égal à 100. La moyenne des méthodes approche 90 et les chiffres confirment bien le fait que l'écart entre la moyenne et la médiane vient d'une distribution faussée du ratio quadratique moyen.

La taille du ratio aux extrêmes de la distribution est trappante dans la distribution du ratio quadratique moyen pour les méthodes A1 et B1. En effet, le maximum atteint 120 dans les deux populations, alors qu'il existe des ratios aussi faibles que 26. Il s'ensuit que l'erreur quadratique moyenne estimée peut augmenter d'un maximum de 20%, mais qu'elle peut aussi baisser de façon assez appréciable pour plusieurs autres estimations. L'ajustement de l'estimation, même

population SDE. Ces figures reposent sur les estimations calculées par Brick et ses collaborateurs (1996) plutôt que celles du tableau 4. La réduction du biais est faible pour les deux groupes. En termes absolus, le biais le plus important s'élève à 1,3% pour la population PE et à 0,9% pour la population SDE. On a calculé la moyenne et la médiane ainsi que la valeur absolue de la réduction du biais pour chaque méthode et chaque groupe. Pour la composante PE, la moyenne et la médiane de la réduction estimée du biais (en termes absolus) se situent entre 0,2 et 0,4% pour les quatre méthodes. Pour la composante SDE, la moyenne et la médiane des valeurs absolues s'établissent entre 0,1 et 0,3%.

### Ratio du biais

L'ordre de grandeur de la réduction absolue du biais n'est pas une statistique très utile de l'incidence du biais, car elle ne tient pas compte de l'ordre de grandeur de l'erreur d'échantillonnage dans l'estimation. Cochran (1977) analyse l'incidence du biais sur les intervalles de confiance d'après la variation du ratio du biais et de l'erreur d'échantillonnage. Pour chaque méthode, le ratio du biais est exprimé par l'équation:

$$(4) \quad r^a = \frac{b^a}{\text{et}(p_s)}$$

où l'erreur-type de l'estimation-type sert de dénominateur. À mesure que le ratio du biais augmente, la probabilité de couvrir la valeur représentant la population s'écarte significativement de l'intervalle de confiance nominal.

Le tableau 4 donne le ratio du biais pour diverses caractéristiques. On remarquera que dans bon nombre de cas, les paramètres de la population PE donnent un ratio important, même si le ratio moyen et le ratio médian sont presque nuls. Dans près de la moitié des cas, le ratio absolu est supérieur à 0,4. Il s'agit d'un ratio assez important pour faire passer l'intervalle de confiance nominal de 95% à environ 93%. Le ratio du biais est plus faible pour les caractéristiques de la population SDE, 15% seulement de ces dernières présentant un ratio supérieur à 0,4.

### 4.2 Erreur quadratique moyenne

La variance ne permettant pas de mesurer correctement l'erreur des estimations biaisées, on se sert pour cela de l'erreur quadratique moyenne (EQM), c'est-à-dire la somme de la variance et du biais de l'estimation porté au carré.

On peut estimer l'EQM pour la NHES:93 d'après l'estimation type de la variance et la réduction estimée du biais qui précède. L'EQM estimée est exprimée par:

$$(5) \quad \text{EQM}^a = \text{var}(p_s) + b^a$$

où  $p_s$  correspond à la proportion estimée obtenue selon l'approche normale et  $b^a$  est la réduction du biais réalisée avec la méthode *a*. Les estimations du biais dérivées des quatre méthodes d'ajustement étant fortement corrélées, on n'a calculé que l'erreur quadratique moyenne de la méthode A1. Brick et ses collaborateurs (1996) ont montré que les estimations qui viennent d'autres méthodes ont un effet négligeable.



Tableau 4 Réduction estimée du biais et ratio du biais pour certaines caractéristiques de la NHES:93 (fin)

Caractéristique	Estimation type				Réduction estimée du biais				Ratio du biais			
	Estimation		Erreur type		Méthode A1		Méthode A2		Méthode B1		Méthode B2	
	Méthode		Méthode		Méthode		Méthode		Méthode		Méthode	

Situation d'emploi du père	26.8	0.6	-0.2	-0.2	-0.1	-0.2	-0.2	-0.1	-0.2	-0.3	-0.2	-0.3
Pas de père dans le ménage	63.2	0.5	0.6	0.9	0.6	0.8	1.2	1.8	1.2	1.8	1.2	1.6
Travail 35 heures/semaine ou plus	3.1	0.2	-0.2	-0.2	-0.2	-0.2	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0
Travail moins de 35 heures/semaine	2.6	0.2	-0.2	-0.3	-0.2	-0.3	-1.0	-1.5	-1.0	-1.5	-1.0	-1.5
Cherche un emploi	4.3	0.3	-0.1	-0.1	-0.1	-0.1	-0.3	-0.3	-0.3	-0.3	-0.3	-0.3
Ne fait pas parti de la population active	91.2	0.3	-0.1	-0.1	-0.1	-0.1	-0.3	-0.3	-0.3	-0.3	-0.3	-0.3

Administration de l'école	8.8	0.3	0.1	0.1	0.1	0.1	0.1	0.3	0.3	0.3	0.3	0.3
Privée	79.9	0.5	0.1	0.4	0.0	0.2	0.2	0.8	0.0	0.4	0.4	0.4
Oui	20.1	0.5	-0.1	-0.4	0.0	-0.2	-0.2	-0.8	0.0	-0.4	-0.4	-0.4
Non	31.5	0.7	-0.6	-0.8	-0.7	-0.9	-0.9	-1.1	-1.0	-1.3	-1.3	-1.3

Programme de sensibilisation aux drogues	68.5	0.7	0.6	0.8	0.7	0.9	0.9	0.9	1.1	1.0	1.3	1.3
Oui	31.5	0.7	-0.6	-0.8	-0.7	-0.9	-0.9	-1.1	-1.0	-1.3	-1.3	-1.3
ou à l'alcool durant l'année	22.3	0.5	-0.3	-0.4	-0.3	-0.5	-0.6	-0.8	-0.6	-1.0	-1.0	-1.0
Non	77.7	0.5	0.3	0.4	0.3	0.5	0.6	0.8	0.6	0.6	0.6	0.6

Lutte de gangs à l'école <sup>6</sup>	22.3	0.5	-0.3	-0.4	-0.3	-0.5	-0.6	-0.8	-0.6	-1.0	-1.0	-1.0
Oui	39.2	0.6	-0.2	-0.3	-0.2	-0.3	-0.3	-0.5	-0.3	-0.5	-0.5	-0.5
Très ou assez facile	29.7	0.5	0.1	0.1	0.2	0.2	0.2	0.2	0.2	0.4	0.4	0.4
Difficile	31.1	0.6	0.1	0.1	0.0	0.1	0.2	0.2	0.2	0.0	0.2	0.2
Presque impossible	66.1	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Crainte de la criminalité à l'école	11.9	0.5	-0.1	-0.2	0.0	-0.2	-0.2	-0.4	0.0	-0.4	0.0	0.0
Aucune	8.6	0.3	-0.1	-0.1	-0.1	-0.1	-0.3	-0.3	-0.3	-0.3	-0.3	-0.3
Crainte de vol ordinaire ou à main armée <sup>5</sup>	13.3	0.5	0.1	0.3	0.1	0.2	0.2	0.6	0.2	0.2	0.4	0.4
Crainte de deux ou plusieurs types d'incidents <sup>5</sup>	38.7	0.6	0.2	0.3	0.1	0.1	0.3	0.6	0.3	0.3	0.6	0.6

Connaissance de la criminalité à l'école	14.1	0.5	0.2	0.3	0.2	0.3	0.4	0.6	0.2	0.2	0.4	0.4
Aucune	15.6	0.4	-0.5	-0.4	-0.4	-0.4	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0
Crainte d'intimidation ou d'agression <sup>5</sup>	31.6	0.6	0.1	0.4	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0
Crainte de deux ou plusieurs types d'incidents <sup>5</sup>	73.0	0.5	0.3	0.2	0.3	0.2	0.6	0.4	0.4	0.6	0.6	0.6

Victime d'un acte criminel	10.9	0.3	-0.2	-0.1	-0.1	0.0	-0.7	-0.3	-0.3	-0.3	-0.3	-0.3
Non	8.9	0.3	-0.1	0.0	0.0	-0.1	-0.3	0.0	0.0	0.0	0.0	0.0
Vol ordinaire ou à main armée <sup>5</sup>	7.2	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Deux ou plusieurs types d'incidents <sup>5</sup>	63.8	0.8	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2

Témoin d'un crime à l'école	24.1	0.8	-0.3	-0.3	-0.3	-0.3	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4
Non	11.4	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Vol à main armée <sup>6</sup>	63.8	0.6	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Intimidation ou agression <sup>6</sup>	24.1	0.8	-0.3	-0.3	-0.3	-0.3	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4

Deux ou plusieurs types d'incidents	63.8	0.6	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Intimidation ou agression <sup>6</sup>	24.1	0.8	-0.3	-0.3	-0.3	-0.3	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4
Vol à main armée <sup>6</sup>	11.4	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Deux ou plusieurs types d'incidents	63.8	0.6	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Source: U.S. Department of Education, National Center for Education Statistics, National Household Education Survey, printemps 1993.

Note: La somme des pourcentages pourrait ne pas correspondre à 100, les chiffres ayant été arrondis.



Tableau 4 Réduction estimée du biais et ratio du biais pour certaines caractéristiques de la NHES:93

Caractéristique	Estimation type			Réduction estimée du biais			Ratio du biais		
	Estimation			Méthode			Méthode		
	Erreur-	A1	A2	B1	B2	A1	A2	B1	B2

Groupe préparation à l'école (PE)	Scolarité des parents	8.6	0.3	-1.7	-1.9	0.1	-5.7	-6.3	0.3
	Diplôme d'études secondaires ou équivalent	33.9	0.8	0.4	0.3	-0.7	0.5	0.4	-0.9
	Etudes collégiales incomplètes	57.5	0.7	1.3	1.6	0.6	1.9	2.3	0.9
	Situation d'emploi de la mère	2.4	0.2	-0.1	-0.1	-0.1	-0.5	-0.5	-0.5
	Pas de mère dans le ménage	34.3	0.5	0.5	0.8	0.2	0.5	1.6	0.4
	Travail 35 heures/semaine ou plus	20.9	0.5	-0.1	-0.2	0.0	-0.2	-0.4	0.0
	Travail moins de 35 heures/semaine	6.6	0.4	0.0	-0.1	-0.1	0.0	-0.3	-0.3
	Cherche un emploi	35.8	0.6	-0.4	-0.3	0.0	-0.7	-0.5	0.0
	Ne fait pas partie de la population active	35.8	0.6	-0.4	-0.3	0.0	-0.7	-0.5	0.0
	Situation d'emploi du père	26.3	0.5	-0.4	-0.6	0.0	-0.1	-1.2	0.0
Groupe sécurité et discipline à l'école (SDE)	Pas de père dans le ménage	63.4	0.6	0.3	0.5	0.1	0.5	0.8	0.2
	Travail 35 heures/semaine ou plus	3.8	0.3	0.0	-0.1	0.0	0.0	-0.3	0.0
	Travail moins de 35 heures/semaine	3.2	0.3	0.0	0.0	-0.1	0.0	0.0	-0.3
	Cherche un emploi	3.3	0.2	0.1	0.2	0.0	0.1	1.0	0.0
	Laps de temps depuis la dernière consultation médicale de routine	84.1	0.4	0.4	0.4	0.2	1.0	1.0	0.5
	Moins d'un an	15.9	0.4	-0.4	-0.5	-0.1	-1.3	-0.5	-0.2
	Poids de naissance	93.3	0.3	-0.1	0.0	0.0	-0.3	0.0	0.3
	5.5 livres ou moins	6.7	0.3	0.1	0.0	0.1	0.0	0.0	0.0
	Plus de 5.5 livres	52.6	0.8	0.9	0.3	0.8	0.6	1.1	0.4
	Enfant inscrit au programme d'un centre <sup>1</sup>	47.4	0.8	-0.9	-0.3	-0.8	-1.1	-0.4	-1.0
Enfant inscrit au programme d'un centre <sup>1</sup>	Oui	52.6	0.8	0.9	0.3	0.8	0.6	1.1	0.4
	Non	47.4	0.8	-0.9	-0.3	-0.8	-1.1	-0.4	-1.0
	Enfant déjà inscrit au programme d'un centre <sup>1</sup>	62.9	0.8	0.5	0.3	0.4	0.6	0.4	0.5
	Oui	37.1	0.8	-0.5	-0.3	-0.4	-0.3	-0.4	-0.5
	Non	37.1	0.8	-0.5	-0.3	-0.4	-0.3	-0.4	-0.5
	A assisté au programme d'un centre avant d'entrer à l'école <sup>2</sup>	73.5	0.5	0.6	0.7	0.5	1.2	1.4	1.0
	Oui	26.5	0.5	-0.6	-0.7	-0.5	-0.6	-1.4	-1.0
	Non	26.5	0.5	-0.6	-0.7	-0.5	-0.6	-1.4	-1.0
	Programme fermes, nourissons et enfants <sup>1</sup>	33.8	1.0	-0.6	-0.1	-0.8	-0.6	-0.1	-0.8
	Oui	66.2	1.0	0.6	0.1	0.8	0.7	0.1	0.8
Repas à l'école ou dans un centre <sup>2</sup>	Oui	35.8	0.6	-0.9	-1.1	-0.5	-1.5	-1.8	-0.8
	Non	64.2	0.6	0.9	1.1	0.5	1.5	1.8	0.8
	Reprise de la maternelle <sup>3</sup>	5.7	0.4	-0.3	-0.5	-0.2	-0.8	-1.3	-0.5
	Oui	94.3	0.4	0.3	0.5	0.2	0.7	1.3	0.5
	Non	94.3	0.4	0.3	0.5	0.2	0.7	1.3	0.5
	Scolarité des parents	9.4	0.5	-1.2	-1.3	-0.3	-0.6	-2.4	-0.6
	Etudes secondaires incomplètes	32.7	0.6	0.3	0.0	-0.2	0.5	0.0	-0.3
	Diplôme d'études secondaires ou équivalent	57.9	0.5	0.9	1.3	0.5	1.1	2.6	1.0
	Etudes collégiales incomplètes	57.9	0.5	0.9	1.3	0.5	1.1	2.6	1.0
	Situation d'emploi de la mère	3.5	0.2	0.0	0.0	0.0	0.0	0.0	0.0
Groupe sécurité et discipline à l'école (SDE)	Pas de mère dans le ménage	46.2	0.5	0.0	0.1	-0.1	0.1	0.2	-0.2
	Travail 35 heures/semaine ou plus	20.3	0.5	0.1	0.0	0.0	-0.1	0.0	0.0
	Travail moins de 35 heures/semaine	4.5	0.3	0.2	-0.2	-0.2	-0.2	-0.7	-0.7
	Cherche un emploi	25.5	0.5	0.0	0.1	0.2	0.2	0.4	0.4
	Ne fait pas partie de la population active	25.5	0.5	0.0	0.1	0.2	0.2	0.4	0.4
Enfant inscrit au programme d'un centre <sup>1</sup>	Pas de père dans le ménage	26.3	0.5	-0.4	-0.6	0.0	-0.1	-1.2	0.0
	Travail 35 heures/semaine ou plus	63.4	0.6	0.3	0.5	0.1	0.5	0.8	0.2
	Travail moins de 35 heures/semaine	3.8	0.3	0.0	-0.1	0.0	0.0	-0.3	0.0
	Cherche un emploi	3.2	0.3	0.0	0.0	-0.1	0.0	0.0	-0.3
	Ne fait pas partie de la population active	3.3	0.2	0.1	0.2	0.0	0.1	1.0	0.0
	Laps de temps depuis la dernière consultation médicale de routine	84.1	0.4	0.4	0.4	0.2	1.0	1.0	0.5
	Moins d'un an	15.9	0.4	-0.4	-0.5	-0.1	-1.3	-0.5	-0.2
	Poids de naissance	93.3	0.3	-0.1	0.0	0.0	-0.3	0.0	0.3
	5.5 livres ou moins	6.7	0.3	0.1	0.0	0.1	0.0	0.0	0.0
	Plus de 5.5 livres	52.6	0.8	0.9	0.3	0.8	0.6	1.1	0.4

#### 4.1 Réduction du biais de couverture

Si on possédait une estimation des mêmes caractéristiques pendant que celles tirées de la NHES:93 mais d'une source indépendante et si ces estimations repères ne souffraient d'aucun biais pour la couverture téléphonique, les cinq estimations pourraient être comparées à leur valeur de référence. Malheureusement, on ne dispose pas de valeurs de référence pouvant servir de point de comparaison avec les estimations des deux volets de la NHES:93. C'est pourquoi, il faut recourir à d'autres méthodes afin d'évaluer les possibilités de réduction du biais par l'ajustement de la couverture.

Faute de valeur de référence, on doit formuler certaines hypothèses en vue d'apprécier l'efficacité des ajustements. Dans le cadre de cette évaluation, on suppose que les méthodes d'ajustement réduisent le biais de couverture. Consécutivement à une telle hypothèse, l'écart entre l'estimation-type et l'estimation ajustée est considéré comme une estimation non biaisée de la réduction du biais de couverture provenant de l'application des méthodes examinées. Mianifestement, les méthodes d'ajustement envisagées ne suppriment pas totalement le biais de couverture. Même si le modèle était exact, la réduction du biais qui affecte les données resterait sujet à l'erreur d'échantillonnage. Malgré les carences qu'elle présente, l'hypothèse doit être formulée si on veut l'utilité de l'ajustement. Si l'ajustement permet d'éliminer le biais, l'erreur quadratique moyenne des estimations après ajustement, équivalra à la variance des estimations, en l'absence de toute contribution du biais de couverture. L'hypothèse modèle penche donc en faveur de l'ajustement des estimations en postulant que les estimations ajustées ne sont pas biaisées. Nous analyserons de façon critique les implications d'une telle hypothèse après avoir fait la preuve de l'efficacité d'une telle méthode.

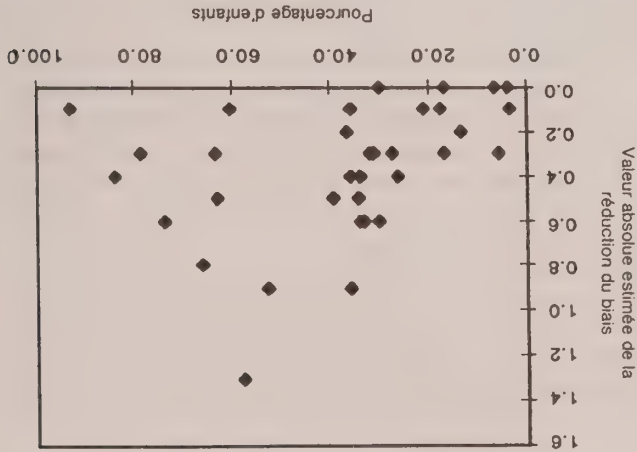
On peut comparer l'estimation de chaque méthode d'ajustement à l'estimation-type de la NHES:93; la différence entre l'estimation-type et l'estimation ajustée donne une idée de la réduction du biais de couverture. Puisqu'il existe quatre estimations ajustées, on peut obtenir quatre estimations différentes de la réduction du biais. La réduction estimée du biais est donnée par:

$$b_a = \hat{p}_s - \hat{p}_a, \quad (3)$$

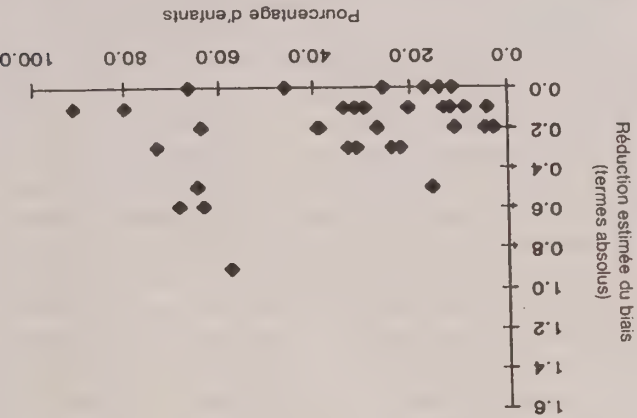
où  $b_a$  représente la réduction estimée du biais obtenue avec la méthode d'ajustement  $a$  ( $a = A1, A2, B1$  ou  $B2$ );  $\hat{p}_s$  est la proportion estimée avec l'estimation-type et  $\hat{p}_a$  est la proportion estimée avec la méthode d'ajustement  $a$ .

Le tableau 4 présente la réduction estimée du biais pour chaque méthode d'ajustement de la pondération. Brick et ses collaborateurs (1996) donnent une estimation analogue pour d'autres caractéristiques. La réduction du biais de l'estimation-type suppose que chaque méthode d'ajustement élimine le biais de couverture.

D'après le tableau 4, on constate que la réduction estimée du biais est inférieure à 1% pour la plupart des aspects et que le sens est cohérent d'une méthode à l'autre. Avant de résumer les estimations, il convient de souligner que le nombre total d'enfants est constant dans les estimations à cause de



**Figure 2.** Valeur absolue de la réduction estimée du biais pour les caractéristiques du groupe «Préparation à l'école» (méthode A1)  
Source: U.S. Department of Education, National Center for Education Statistics, National Household Education Survey, printemps 1993.



**Figure 3.** Valeur absolue de la réduction estimée du biais pour les caractéristiques du groupe «Sécurité et discipline à l'école» (méthode A1)  
Source: U.S. Department of Education, National Center for Education Statistics, National Household Education Survey, printemps 1993.

L'itération des estimations par rapport aux totaux de la CPS. Le fait que le nombre total d'enfants soit fixe dans les catégories de réponse a deux conséquences: il s'ensuit une correction négative de la réduction estimée du biais d'une catégorie de réponses à l'autre et on obtient une fausse impression du nombre d'éléments d'information indépendants dans les valeurs du tableau.

Pour résoudre le problème que soulève la réduction des estimations du biais, on a décidé de supprimer l'estimation d'une des catégories de réponse pour chaque élément. Ainsi, on a éliminé la réponse «non» dans toutes les catégories pour lesquelles la réponse pouvait être «oui» ou «non». Pour les variables d'un autre genre, on a choisi la catégorie de réponse qui présentait l'estimation la plus faible.

La figure 2 illustre la valeur absolue de la réduction du biais estimée avec la méthode A1 pour les caractéristiques de la population PE. La figure 3 illustre la même chose pour la



(1) 
$$w'_i = w_i \left( 1 + \delta_i \frac{\frac{t_1 + t_2}{t_4}}{\frac{t_1^* + t_2^*}{t_2^*}} \right)$$

où  $w_i$  est le poids de la NHES:93 ajusté pour tenir compte de la non-réponse des personnes échantillonnées, sans itération par rapport aux totaux d'octobre 1992 de la CPS, et où  $\delta_i = 1$  quand la personne fait partie d'un ménage où le service téléphonique a été interrompu durant l'année antérieure et est égal à zéro dans les autres cas. La quantité entre parenthèses de l'équation (1) est une estimation de A, le coefficient d'ajustement des poids.

On a calculé séparément les nouveaux poids des populations PE et SDE. Au lieu de procéder à l'ajustement global tel qu'indiqué en (1), on a mesuré la pondération dans les cellules définies par chacune des quatre méthodes d'ajustement (A1, A2, B1 et B2). Le tableau 2 donne les coefficients d'ajustement résultant pour les populations PE et SDE. Les ajustements qui apparaissent à la première colonne correspondent à ceux des méthodes A1 et B1. Ceux de la deuxième colonne résultent des facteurs des méthodes A2 et B2. Les coefficients d'ajustement des méthodes qui supposent une interruption du service téléphonique d'un mois ou davantage sont plus importants que ceux des méthodes présupposant une interruption d'une semaine ou plus, le dénormalisateur du quotient étant par définition plus petit avec cette classification (voir la figure 1 pour l'estimation du pourcentage de personnes qui ont connu une interruption de service dans chaque cas). La dernière étape de la pondération implique l'itération des quatre poids possibles aux totaux d'octobre 1992 de la CPS utilisées pour l'itération des poids types de la NHES:93 s'appliquant aux personnes. Cette méthode débouche sur le poids type de la NHES:93 et quatre autres poids dérivant des méthodes d'ajustement. Les cinq poids se conformant aux mêmes totaux marginaux. La seule différence concerne l'ajustement relatif à l'interruption du service téléphonique effectué avant l'itération. Les poids types ne sont pas ajustés davantage, mais les autres poids sont ajustés différemment, selon la méthode retenue.

Tableau 3

Ratio du facteur d'inflation de la variance attribuable à l'ajustement de la couverture

Groupe	Taille de l'échantillon	FIV*	Ratio du facteur d'inflation de la variance attribuable à l'ajustement de la couverture
Préparation à l'école	10,888	1.36	1.20
Sécurité et discipline à l'école	2,563	1.37	1.12
3 <sup>e</sup> à 5 <sup>e</sup> année	10,117	1.39	1.13
6 <sup>e</sup> à 12 <sup>e</sup> année	12,680	1.49	1.12
3 <sup>e</sup> à 12 <sup>e</sup> année			
			1.26
			1.24
			1.11
			1.25

\* L'acronyme FIV désigne le facteur d'inflation type. Il s'agit du coefficient de variation des poids porté au carré, plus un. Source: U.S. Department of Education, National Center for Education Statistics, National Household Education Survey, printemps 1993.

4. CONSTATIONS

Comme on peut le voir ci-dessus, en ajustant les poids pour atténuer le biais, on en accroît la variabilité. Par conséquent, la variance des estimations augmente. Kish (1992) propose une expression pour mesurer approximativement la hausse de la variance attribuable à l'application de poids inégaux. Cette expression est baptisée facteur d'inflation de la variance (FIV). Le FIV prend la forme suivante:

(2) 
$$FIV = 1 + CV^2(\text{poids})$$

où CV représente le coefficient de variation des poids.

Le tableau 3 donne le FIV des poids types de la NHES:93 pour chaque groupe. Le groupe SDE est divisé en fonction de la scolarité, les enfants ayant été sélectionnés à des taux différents selon l'année d'étude. Le FIV de chaque groupe est environ 1.4. En d'autres termes, la variance a augmenté d'environ 40% à cause de la variabilité des poids types. Le FIV du fichier SDE dans son ensemble est un peu plus élevé (1.5) puisqu'il inclut des jeunes échantillonnés à des taux différents.

Les autres facteurs du tableau 3 correspondent au ratio du FIV des quatre autres poids et du FIV du poids type. Le résultat indique dans quelle mesure la variance devrait augmenter quand on utilise les autres poids au lieu des poids types de la NHES:93. Dans l'ensemble, la hausse de la variance attribuable à l'ajustement de la couverture d'après l'interruption du service téléphonique varie de 9 à 13% pour les méthodes A1 et B1 appliquées à la population SDE, mais elle peut monter jusqu'à 20% pour la population PE. Les ratios sont plus élevés pour les méthodes A2 et B2 puisqu'ils se situent entre 24 et 35%. Le ratio le plus important correspond à celui de la méthode A2 appliquée à la population PE. Les ratios (et le FIV, par voie de conséquence) plus élevés obtenus avec les méthodes reposant sur les interruptions de service d'un mois ou davantage sont attribuables aux coefficients plus importants et plus variables qui apparaissent à la deuxième colonne du tableau 2. Les ratios de la population PE sont plus élevés que ceux de la population SDE.



sur la scolarité des parents ou la propriété (d'une habitation). On s'est servi de la race pour créer des cellules à l'intérieur des catégories «scolarité» et «propriété». Ces cellules ont été retenues parce que le pourcentage de personnes rapportant une interruption du service téléphonique varie avec ces caractéristiques, et parce que les données correspondantes étaient disponibles pour la CPS. De là, on a défini quatre méthodes d'ajustement:

**Méthode A1** – enfants des ménages où le service téléphonique a été interrompu pendant au moins une semaine, selon la scolarité des parents (études secondaires incomplètes, diplôme d'études secondaires, études collégiales ou études supérieures) et la race (hispanique, noir/non hispanique, blanc et autres/non hispanique);

**Méthode A2** – enfants des ménages où le service téléphonique a été interrompu un mois ou davantage, selon la scolarité des parents et la race;

**Méthode B1** – enfants des ménages où le service téléphonique a été interrompu au moins une semaine, selon la propriété (propriétaire/autre, locataire) et la race;

**Méthode B2** – enfants des ménages où le service téléphonique a été interrompu un mois ou davantage, selon la propriété et la race.

On n'a pas pu obtenir les facteurs d'ajustement des méthodes qui précèdent directement des données de la NHES:93, faute de données sur les ménages sans téléphone. On a donc procédé aux ajustements en prenant les données de la CPS et de la NHES:93, puis en les appliquant aux poids de la NHES:93.

Tableau 2

Coefficients d'ajustement pour la pondération des cellules, selon la durée de l'interruption du service téléphonique

Facteur	Durée de l'interruption			
	SDE	PE	Une semaine ou plus	Une mois ou plus

Cellules définies par la scolarité des parents et la race (méthodes A1 et A2)

Études secondaires incomplètes; hispanique 5.75  
Études secondaires incomplètes; noir, non hispanique 5.10  
Études secondaires incomplètes; blanc et autre, non hispanique 4.98  
Diplôme d'études secondaires; hispanique 2.31  
Diplôme d'études secondaires; noir, non hispanique 2.65  
Diplôme d'études secondaires; blanc et autre, non hispanique 2.16  
Études collégiales ou supérieures; hispanique 1.34  
Études collégiales ou supérieures; noir, non hispanique 1.77  
Études collégiales ou supérieures; blanc et autre, non hispanique 1.58

Cellules définies par la propriété et la race (méthodes B1 et B2)

Locataire; hispanique 3.74  
Locataire; noir, non hispanique 3.23  
Locataire; blanc et autre, non hispanique 2.43  
Propriétaire/autre; hispanique 2.00  
Propriétaire/autre; noir, non hispanique 2.53  
Propriétaire/autre; blanc et autre, non hispanique 2.26

Cellules définies par la scolarité des parents et la race (méthodes A1 et A2)

Études secondaires incomplètes; hispanique 4.89  
Études secondaires incomplètes; noir, non hispanique 4.26  
Études secondaires incomplètes; blanc et autre, non hispanique 3.81  
Diplôme d'études secondaires; hispanique 2.67  
Diplôme d'études secondaires; noir, non hispanique 3.06  
Diplôme d'études secondaires; blanc et autre, non hispanique 2.18  
Études collégiales ou supérieures; hispanique 1.96  
Études collégiales ou supérieures; noir, non hispanique 1.35  
Études collégiales ou supérieures; blanc et autre, non hispanique 1.91

Cellules définies par la propriété et la race (méthodes B1 et B2)

Locataire; hispanique 3.58  
Locataire; noir, non hispanique 3.38  
Locataire; blanc et autre, non hispanique 2.99  
Propriétaire/autre; hispanique 2.81  
Propriétaire/autre; noir, non hispanique 2.90  
Propriétaire/autre; blanc et autre, non hispanique 2.03

à un programme d'assistance publique (« Femmes, nourrissons et enfants » ou repas gratuits) se caractérisent par un taux d'interruption du service téléphonique beaucoup plus élevé que les non-participants à un tel programme. Le pourcentage d'enfants des ménages qui connaissent des interruptions du service téléphonique varie toutefois moins avec les caractéristiques associées à la préparation à l'école ainsi qu'à la sécurité et à la discipline à l'école qu'avec les paramètres socioéconomiques. Brick, Keefer, Waksberg et Bell (1996) ont examiné d'autres caractéristiques des deux populations, mais nous ne les reproduisons pas ici. Pour la plupart des autres grands paramètres, la variation du pourcentage de personnes ayant connu une interruption quelconque du service téléphonique n'est pas statistiquement significative, ou n'est pas assez importante pour présenter une véritable importance pratique.

### 3. AJUSTEMENT DES POIDS

Presque toutes les enquêtes par échantillonnage ajustent les données des répondants pour tenir compte des non-réponses et de la non-couverture, et ainsi atténuer la variabilité des estimations au moyen de données auxiliaires issues d'autres sources. Un des principaux avantages que présente un tel ajustement, pour les enquêtes téléphoniques par échantillonnage, est que de cette façon, on atténue souvent le biais associé au sous-dénombrement des membres des ménages sans téléphone.

Kalton et Kasprzyk (1986) parlent des ajustements aux poids de base et les classent en quatre catégories: ajustement des poids de la population, ajustement des poids de l'échantillon, méthode itérative du quotient et ajustement de la probabilité de réponse. La NHES:93 procédait à un ajustement du poids de l'échantillon et à l'application de la méthode itérative du quotient pour tenir compte d'une variation de la non-réponse des personnes échantillonnées. La méthode itérative du quotient a ensuite permis de faire concorder les distributions marginales de l'échantillon avec les totaux de la Current Population Survey (CPS) d'octobre 1992. Un des principaux avantages de la méthode itérative du quotient est que la non-réponse des personnes avec et sans téléphone.

Bien que les pondérations puissent atténuer le biais attribuable au sous-dénombrement, pareils ajustements ont aussi généralement pour effet d'augmenter la variance des estimations. Kish (1992) parle des raisons pour lesquelles on se sert de poids inégaux et des conséquences que cela entraîne dans diverses situations. Il préconise une approche statistique double qui consiste à équilibrer la réduction du biais et la hausse de la variance. Si les poids atteignent sensiblement le biais des estimations, il pourrait valoir la peine de laisser la variance augmenter. Une faible réduction du biais associée à un net relèvement de la variance, en revanche, n'est pas recommandée.

Le reste de la présente partie traite des méthodes spécifiques d'ajustement des poids. On y trouvera les propriétés statistiques des poids élaborés selon quatre méthodes d'ajustement différentes. Les autres poids sont appliqués aux données de la NHES:93, et on compare la réduction du biais des estimations à la hausse de la variance attribuable à la pondération inégale.

#### Méthodes d'ajustement

En premier lieu, il convient de décider comment on classera la durée de l'interruption du service téléphonique. On a examiné des interruptions de longueur variable pour déterminer à quel point les interruptions temporaires sans raisons économiques peuvent être différenciées des autres. On a finalement convenu d'articuler les cellules d'ajustement sur deux types d'interruption: celles d'une semaine ou plus et celles d'au moins un mois.

Pour chaque catégorie relative à la durée de l'interruption, on a réparti les enfants entre des cellules d'ajustement reposant



**Tableau 1**  
Pourcentage estimé de personnes qui ont connu une interruption du service téléphonique au cours des douze mois antérieurs pour trois populations

NSV		NHES:93 (PE)		NHES:93 (SDE)	
Estimation	Erreur-type	Estimation	Erreur-type	Estimation	Erreur-type

Total	2.3	0.1	12.0	0.4	9.2	0.3
Région						
Mid-Ouest	2.3	0.2	11.0	1.0	7.3	0.7
Nord-est	2.0	0.2	9.5	1.2	9.0	0.8
Sud	2.6	0.2	13.6	0.7	10.8	0.6
Ouest	2.4	0.2	12.5	0.9	9.2	0.8
Race/ethnie <sup>1</sup>						
Blanc	2.0	0.1	9.3	0.5	7.2	0.3
Noir	3.5	0.4	19.8	1.5	14.7	1.1
Hispanique	3.9	0.5	17.2	1.5	14.1	1.1
Autre	2.6	0.6	11.7	2.6	9.3	1.5
Scolarité <sup>2</sup>						
Études secondaires incomplètes	3.2	0.2	18.4	1.8	17.4	1.6
Diplôme d'études secondaires	2.0	0.2	15.4	0.8	11.0	0.8
Cours collégial incomplet	2.3	0.2	11.8	0.7	8.6	0.5
Baccalauréat	1.6	0.2	5.5	0.8	5.3	0.8
Études supérieures	2.2	0.3	5.2	0.7	4.5	0.6
Revenu du ménage						
\$10,000 ou moins	22.8		1.3		19.0	1.3
\$10,001 à \$20,000	19.9		1.4		15.7	1.1
\$20,001 à \$30,000	9.3		0.8		7.9	0.6
Plus de \$30,000	5.5		0.5		5.0	0.3
Participant au programme						
Femmes, nourrissons et enfants <sup>3</sup>						
Oui	18.2		1.3			
Non	8.0		0.6			
Repas gratuit à l'école ou dans un centre <sup>4</sup>						
Oui	21.1		1.2			
Non	7.6		0.5			
Poids de naissance						
5.5 livres ou moins	12.0		1.6			
Plus de 5.5 livres	12.0		0.4			
Administration de l'école						
Publique	9.4					
Privée	7.5					
Facilité d'obtenir de la marijuana à l'école <sup>5</sup>						
Très ou assez facile	9.7					
Difficile	8.0					
Presque impossible	9.0					

<sup>1</sup> La race ou l'ethnie s'applique au membre le plus âgé du ménage dans la NSV et à l'enfant dans la NHES:93.

<sup>2</sup> La scolarité est celle du membre le plus âgé du ménage dans la NSV et du parent le plus instruit de l'enfant dans la NHES:93.

<sup>3</sup> Estimation restreinte aux enfants d'âge préscolaire.

<sup>4</sup> Estimation applicable aux enfants sauf à ceux d'âge préscolaire.

<sup>5</sup> Estimation applicable uniquement aux enfants de la 6<sup>e</sup> à la 12<sup>e</sup> année.

Source: U.S. Department of Veterans Affairs, National Survey of Veterans, été/automne 1993 et U.S. Department of Education, National Household Education Survey, printemps 1993.



La variation des estimations attribuable à la formulation de la question sur l'interruption du service téléphonique est manifeste dans les résultats de deux enquêtes effectuées en Virginie par la Virginia Community University. Lors de l'enquête de novembre 1993, on s'est enquis des interruptions du service téléphonique en représentant le libellé de la NSV, en avril 1994 cependant, on s'est servi de la question de la NHES:93. Les résultats reproduisent l'écart noté entre les estimations de la NSV et de la NHES:93. Ainsi, l'enquête de novembre 1993 situe la proportion de ménages dont le service téléphonique avait été interrompu au cours des 12 mois antérieurs à 3%, alors que l'enquête d'avril l'établit autour de 9%. On en conclut que la manière de poser la question influe lourdement sur l'ordre de grandeur de l'estimation, et il est plausible que les estimations de la NSV soient biaisées vers le bas. Certains adultes dont le service téléphonique a bel et bien été interrompu au cours des 12 mois antérieurs ont vraisemblablement mal répondu à la question de la NSV.

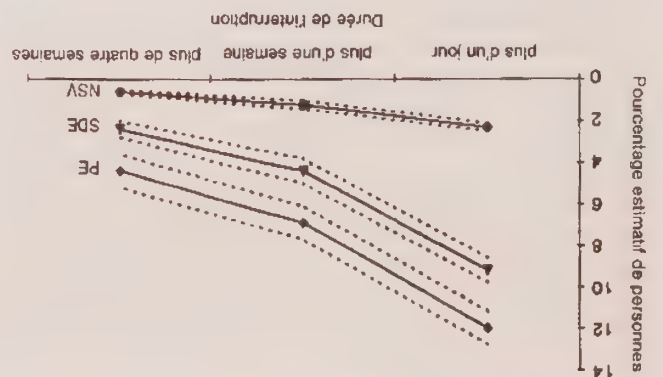
### Caractéristiques des personnes qui ont connu une interruption de service

Le pourcentage estimé de personnes dont le service téléphonique a été interrompu est examiné ci-dessous en fonction des caractéristiques de ces personnes. L'exercice a pour but d'établir si les données pourraient servir à corriger le biais de couverture. Nous avons estimé le pourcentage de personnes dans les ménages où le service téléphonique avait été interrompu d'après les caractéristiques recueillies dans le cadre de la NSV et de la NHES:93. Ces estimations apparaissent à la première partie du tableau 1. Il se peut que certains écarts dans la distribution résultent de la façon dont les questions ont été posées. Ainsi, les deux enquêtes ne classent pas la scolarité de la même façon: dans la NSV, on note le niveau de scolarité de la personne la plus âgée du ménage alors que dans la NHES:93, on retient le niveau de scolarité du plus instruit des deux parents de l'enfant.

Le reste de notre analyse se restreint aux données de la NHES:93 pour deux raisons. Tout d'abord, l'interview détaillé des volets PE et SDE de la NHES:93 nous renseigne plus sur les caractéristiques des ménages que l'interview préliminaire de la NSV. Deuxièmement, les interruptions du service téléphonique estimées par la NSV sont biaisées à cause de la façon dont la question est formulée, ainsi que nous l'avons expliqué plus haut. Bien sûr, les estimations de la NHES:93 concernent des ménages avec enfants, pour qui le taux de non-abonnement au service téléphonique est plus élevé que dans la population en général. En ce sens, elles ne reflètent donc pas la réalité.

Les données de la NHES:93 révèlent que le pourcentage de personnes qui connaissent une interruption quelconque du service téléphonique est relativement cohérent pour les populations PE et SDE (voir le tableau 1). Le pourcentage d'interruptions le plus élevé est lié aux caractéristiques habituellement associées à une situation économique difficile. Le pourcentage d'enfants des populations PE et SDE qui connaissent une interruption du service téléphonique est plus élevé dans les ménages à faible revenu que dans les ménages mieux nantis. De même, les enfants inscrits

Figure 1. Pourcentage estimé de personnes dont le service téléphonique a été interrompu dans les trois populations



L'écart important entre les estimations de la NSV et de la NHES:93 est attribuable à au moins deux grandes raisons. Tout d'abord, il s'agit de populations différentes. Ainsi, il est prévisible que les enfants en bas âge vivent dans un ménage où les interruptions de service sont plus fréquentes que dans un ménage composé d'adultes et d'enfants plus âgés. Thornberry et Massey (1988) estiment que le taux de couverture de l'enquête téléphonique est plus faible pour les enfants en bas âge que pour n'importe quel autre groupe d'âge. L'écart d'environ 3% dans l'estimation du pourcentage de personnes qui ont connu une interruption du service téléphonique observé entre le groupe des enfants en bas âge (PE) et celui des enfants plus vieux (SDE) de la NHES:93 est donc raisonnable.

Des populations différentes n'expliquent toutefois pas entièrement l'importante variation entre les estimations de la NSV et de la NHES:93. Une des principales raisons à cela tient à la façon dont les questions ont été posées dans les deux enquêtes. L'interview de la NHES:93 débutait par une question du genre: «Au cours des 12 derniers mois, votre ménage a-t-il connu une interruption du service téléphonique de plus de 24 heures?». Lors de l'interview de la NSV, on avait plutôt demandé: «Votre ménage n'a-t-il pas été sans service téléphonique pendant au moins 24 heures?». La NSV recourait donc à une question de sélection complétée par une question plus détaillée. Une construction de ce genre débouche souvent sur une sous-estimation de l'activité à laquelle on s'intéresse, ce que semblait confirmer les résultats plus faibles de la NSV. Une raison plus importante expliquant la différence observée a sans doute trait au libellé des questions. Avec la question de la NSV, la réponse «non» aurait pu sembler la confusion chez le répondant, puisqu'on lui demandait s'il n'avait pas été sans service téléphonique. Converse et Presser (1986) expliquent des difficultés d'une telle formulation. Le libellé prête moins à confusion dans la NHES:93. La façon dont la question est formulée et le recours à une question de sélection dans la NSV pourraient bien être essentiellement à l'origine de la plus faible estimation obtenue avec ce questionnaire.



Lors de l'interview préliminaire, les membres du ménage de 14 ans et plus ont été recensés, puis on leur a posé des questions afin d'établir leurs paramètres et leur condition d'ancien combattant. L'interview plus détaillée a été adm-nistrée aux adultes qui avaient déclaré être un ancien combattant. Les résultats présentés plus loin s'appliquent aux adultes recensés durant l'enquête préliminaire, laquelle n'avait d'autre but qu'établir quelques caractéristiques des adultes et du ménage.

Lors de la NHES:93, 64,000 ménages ont été passés en revue et on a procédé à près de 30,000 interviews parmi les ménages échantillonnés. L'enquête comportait deux volets: la préparation à l'école (PE), et la sécurité et la discipline à l'école (SDE). Environ 11,000 parents d'enfants de 3 à 7 ans ont répondu à l'interview sur la PE et approximativement 12,700 parents d'enfants de la troisième à la douzième année, aux éléments du volet SDE. On a recueilli des renseignements sur les interruptions du service téléphonique auprès des ménages qui avaient subi au moins une entrevue complète sur la PE ou la SDE.

Les réponses aux questions de la NHES:93 ne venant que des ménages qui avaient entièrement répondu à l'interview sur la PE ou la SDE, il est possible d'analyser de nombreuses caractéristiques des enfants, mais les données ne couvrent pas une population aussi vaste que celle de la NSV. D'un autre côté, si la NSV touchait l'ensemble des adultes, on ne possède que des données fragmentaires sur la majorité d'entre eux. On a demandé à un membre des ménages qui avaient été interviewés (interview préliminaire de la NSV et interview plus détaillée de la NHES:93) si le service téléphonique avait été interrompu au cours des 12 mois précédents et qu'elle avait été la durée de l'interruption.

## Interruption estimée du service téléphonique selon la NSV et la NHES:93

Le pourcentage estimé de personnes d'un ménage qui avaient interrompu le service téléphonique au moins une journée au cours des 12 mois antérieurs varie sensiblement d'une enquête à l'autre. En effet, 2,3% seulement des adultes avaient connu une interruption de service d'un jour ou plus, selon les données de la NSV, contre 12,0% pour le volet de la NHES:93 sur les enfants en bas âge (population PE de 3 à 7 ans) et 9,2% pour la population SDE d'enfants plus âgés (3<sup>e</sup> à 12<sup>e</sup> année).

La figure 1 montre l'estimation et l'intervalle de confiance à 95% du pourcentage de personnes dont le service téléphonique a été interrompu au moins un jour, plus le pourcentage estimé de personnes qui ont connu une interruption de service d'au moins une semaine et de quatre semaines ou davantage. Quoique le pourcentage fluctue d'échantillon en échantillon, les tendances de la hausse selon la durée de l'interruption sont relativement stables. Le pourcentage de ménages qui connaissent une interruption de service d'au moins une semaine correspond à moins de la moitié du pourcentage de ménages dont le service est interrompu pour une durée quelconque, et la proportion de ménages sans interruption pendant quatre semaines et plus représente environ le quart du pourcentage de ménages qui ont connu une interruption quelconque du service.

Une autre grande condition à satisfaire pour qu'on puisse se servir des ménages transitoires afin d'atténuer le biais de couverture concerne les caractéristiques de ces ménages et des ménages sans téléphone. Si les deux groupes diffèrent, l'ajustement manquera d'efficacité. Se servant des données du panel et des données de plusieurs enquêtes effectuées en Virginie, Keeter (1995) a montré que les caractéristiques des ménages transitoires se rapprochent beaucoup plus de celles des ménages sans téléphone que de celles des ménages abonnés.

Ces constatations nous amènent à croire que les facteurs de pondération reposant sur les données issues des ménages abonnés au service téléphonique une partie de l'année seulement donneraient de meilleurs résultats que les facteurs de pondération actuellement utilisés. Pour évaluer cette méthode d'ajustement des poids, on a ajouté des questions à deux enquêtes nationales effectuées par Westat en 1993. Les deux sondages recouraient à la composition aléatoire (CA) et à l'interview téléphonique assistée par ordinateur. Les données ont été recueillies aux centres de recherches téléphoniques de Westat.

La première enquête était la National Household Education Survey de 1993 (NHES:93). Cette dernière a été entreprise au nom du National Center for Education Statistics du Département de l'éducation, au printemps de 1993, et portait sur les problèmes relatifs, d'une part, à la préparation des enfants en bas âge à l'école, et d'autre part, à la sécurité et à la discipline à l'école. L'autre sondage était la National Survey of Veterans (NSV), effectuée au deuxième semestre de 1993 pour le Département des anciens combattants américain. À cette occasion, on a interrogé les adultes pour savoir s'ils étaient des anciens combattants. À ceux qui répondaient par l'affirmative, on a posé des questions sur divers sujets, notamment la santé, l'éducation et la situation financière.

Plus loin, nous présentons une estimation du pourcentage de personnes qui avaient temporairement interrompu leur service téléphonique, nous décrivons comment ajuster les poids du sondage à partir de ces données et nous analysons les conséquences statistiques de l'utilisation des nouveaux facteurs de pondération. La dernière partie du document résume les constatations et contient certains commentaires au sujet de l'application de cette technique au sondage téléphonique CA.

## 2. ESTIMATION DES INTERRUPTIONS DU SERVICE TÉLÉPHONIQUE

Le pourcentage de personnes qui mettent temporairement fin à leur abonnement au service téléphonique doit être estimé avant qu'on puisse déterminer s'il peut réduire le biais de couverture d'un sondage national. Des questions ont été ajoutées à la NSV et à la NHES:93 à cette fin. On a sélectionné environ 23,000 ménages dans le cadre de la NSV, puis interviewé au-delà de 5,500 anciens combattants admissibles.

# Utilisation des données sur les interruptions du service téléphonique pour ajuster la couverture

J. MICHAEL BRICK, JOSEPH WAKSBERG et SCOTT KEETER<sup>1</sup>

## RÉSUMÉ

La couverture des sondages téléphoniques effectués aux États-Unis est biaisée, car environ 6% des ménages sont privés de téléphone à un moment ou à un autre dans le temps. Le biais attribuable au sous-décomptement peut être important. En effet, les ménages sans téléphone sont généralement plus démunis que les autres et leurs caractéristiques diffèrent de celles de la population d'abonnés. La stratification a posteriori et les autres méthodes d'ajustement habituelles permettent rarement de compenser pareil biais en totalité. La présente recherche porte sur une méthode servant à ajuster les estimations de l'enquête. Cette méthode repose sur l'observation que certains ménages n'ont le téléphone qu'une partie de l'année, souvent à cause de difficultés économiques. En recueillant des données sur les interruptions du service téléphonique durant l'année antérieure, on peut procéder à un ajustement statistique des estimations et réduire le biais, mais parallèlement la variance augmente en raison de la plus grande variabilité des poids. Nous examinerons ici une méthode d'ajustement articulée sur les données recueillies lors d'un sondage téléphonique menée à l'échelle nationale. La réduction du biais et l'effet sur l'erreur quadratique moyenne des estimations ont été évalués pour diverses statistiques. Les résultats indiquent que lorsque les estimations tirées de l'enquête sont étroitement liées à la situation économique, la méthode d'ajustement fondée sur les interruptions du service téléphonique peut améliorer l'erreur quadratique moyenne des résultats.

MOTS CLÉS: Couverture; biais; ajustement de la pondération; échantillonnage de numéro de téléphone; enquêtes téléphoniques par composition aléatoire.

## 1. INTRODUCTION

Les sondages téléphoniques permettent de recueillir des données de façon relativement plus économique que l'interview directe. Aux États-Unis, cependant, ces sondages sont teints par une importante source de biais à laquelle échappent les enquêtes-ménages recourant à l'interview directe. En effet, actuellement, 94% seulement des ménages du pays bénéficient du service téléphonique à un moment quelconque dans le temps. D'autre part, le taux de couverture est encore plus faible pour certains groupes comme les ménages avec enfants en bas âge.

Une pondération reposant sur une stratification a posteriori d'après les variables démographiques qu'on sait être associées à la couverture téléphonique atténue en partie les conséquences d'une couverture biaisée lors d'un sondage téléphonique. Toutefois, même lorsqu'elle est efficace, la pondération en fonction des totalisations démographiques connues ne résout pas entièrement le problème de couverture, car la compensation est insuffisante pour certaines variables (Massey et Botman 1988) et exagérée pour d'autres (Brick, Burke et West 1992).

Le présent article porte sur une autre méthode permettant d'ajuster les données des sondages téléphoniques, de manière à tenir compte du biais de couverture. Cette méthode, proposée par Keeter (1995), repose sur l'observation que l'abonnement au service téléphonique varie de façon dynamique non seulement d'un ménage à l'autre, mais aussi d'un moment à l'autre dans le temps pour le même ménage. En effet, un nombre appréciable de ménages américains s'abonnent au

service téléphonique ou annulent leur abonnement pendant l'année. À cause de ce phénomène, la population d'abonnés à un point quelconque dans le temps comprend des ménages qui faisaient partie de la population de non-abonnés peu de temps auparavant. Malgré une somme considérable de renseignements sur la taille et les caractéristiques de la population de non-abonnés, on sait peu de choses sur la dynamique à court terme de cette dernière. Les données émanant des travailleurs sociaux, des compagnies de téléphone et de ceux qui interviennent auprès des ménages en difficulté laissent supposer que l'abonnement au service téléphonique est sporadique dans bon nombre de cas. Un ménage peut s'abonner au service téléphonique quand il est en mesure de le faire et mettre fin à son abonnement lorsque surrissent des difficultés ou quand la facture devient trop lourde à régler (Fédéral Communications Commission 1988). On ignore combien de ménages sautent d'un état à l'autre et le laps de temps pendant lequel ils demeurent dans tel ou tel état.

Keeter (1995) a étudié deux enquêtes par panel de ménages pour estimer la dynamique des abonnements au service téléphonique. Les ménages qui changent d'état (existence du téléphone dans le ménage) sont baptisés ménages «transitoires». Dans le cas d'une enquête par panel dont les données ont été recueillies à douze mois d'intervalle, la moitié des 6% de ménages sans téléphone répondraient à la définition de «transitoires» aux deux dates d'enquête par panel, où les données ont été recueillies à deux mois d'intervalle. Puisque ces estimations reposent sur des observations effectuées à deux points dans le temps plutôt que sur une mesure continue, elles sous-estiment

<sup>1</sup> J. Michael Brick et Joseph Waksberg, Westat, Inc., 1650 Research Blvd., Rockville, MD 20850, U.S.A.; Scott Keeter, Virginia Commonwealth University, Survey Research Laboratory, Richmond, VA 23284, U.S.A.



la méthode AATC, a donné des résultats supérieurs d'au moins dix pour cent au résultat des autres méthodes pour chacune des fonctions  $m_i(z)$ , et a donc été retenu de préférence pour la production des cartes du ABARE.

La méthode de sélection de la largeur de bande visant à réduire l'aspect tacheté d'une carte (AATC) est une mesure du lissage de la carte: elle détermine le degré de similitude de la valeur lissée d'une exploitation quelconque, par rapport à celle de ses voisines. Désignons par  $p(i)$  l'estimation du percentile de la variable lissée à la  $i$ -ième exploitation. Désignons en outre par  $\delta_i$  l'ensemble des indices des six exploitations les plus proches de cette  $i$ -ième exploitation. Cette méthode est fondée sur le calcul de

$$SF(h) = (6n)^{-1} \sum_{k \in \delta_i} |p(i) - p(k)|$$

Cette valeur est indépendante de l'échelle et diminue monotoniqument avec la réduction de la largeur de bande. La largeur de bande choisie est la plus faible qui présente un taux de réduction de AATC ( $< \epsilon$ ) suffisamment faible. La valeur de  $\epsilon$  a été choisie subjectivement à la suite d'un examen détaillé des cartes de cinq variables clés pour cinq valeurs de  $\epsilon$ .

## BIBLIOGRAPHIE

- BANKIER, M.D., RATHWELL, S., et MAJKOWSKI, M. (1992). Two step generalised least squares estimation in the 1991 Canadian Census. *Proceedings of the Workshop on Uses of Auxiliary Information in Surveys*. Statistics Sweden, Örebro, Octobre 5-7.
- BARDSLEY, P., et CHAMBERS, R.L. (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics*, 33, 290-299.
- BRECKLING, J., et CHAMBERS, R.L. (1988). *M-quantiles*. *Biometrika*, 75, 761-771.
- DEVILLE, J.-C., et SÄRNDA, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- FULLER, W.A., LOUGHIN, M.M., et BAKER, H.D. (1994). Production de poids de régression en situation de non-réponse et application à la Nationwide Food Consumption Survey de 1987-1988. *Techniques d'enquête*, 20, 79-89.
- HALL, P. (1992). *The Bootstrap and Edgeworth Expansion*. New York: Springer-Verlag.
- HÄRDLE, W. (1990). *Applied Nonparametric Regression*. Cambridge: Cambridge University Press.
- NADARAYA, E.A. (1964). On estimating regression. *Theory of Probability and its Applications*, 10, 186-190.
- NEWBY, W.K., et POWELL, J.L. (1987). Asymmetric least squares estimation and testing. *Econometrica*, 55, 819-847.
- RUPPERT, D., et WAND, M.P. (1994). Multivariate locally weighted least squares regression. *Annals of Statistics*, 22, à paraître.
- WAND, M.P., et JONES, M.C. (1995). *Kernel Smoothing*. London: Chapman and Hall.
- WATSON, G.S. (1964). Smooth regression analysis. *Sankhyā*, série A, 26, 101-116.

Dans le présent article, nous avons démontré que lorsque des données d'enquête ont une dimension spatiale, comme c'est le cas pour le AAGIS, les notions de lissage spatial peuvent être utiles à l'analyste. On peut y avoir recours pour modifier les poids et réduire ainsi la variabilité des estimations portant sur les petites régions. Il peut également être utile de procéder au lissage des dimensions spatiales des données avant la cartographie de la fonction de la moyenne spatiale.

Comme le présent article porte sur la cartographie des données, nous avons limité notre propos au lissage des dimensions spatiales. Toutefois, les mêmes méthodes peuvent servir au lissage en fonction d'autres dimensions. Ainsi, si on a des raisons de s'attendre à une forte corrélation sérielle lorsque la population sous-jacente est ordonnée en fonction d'une variable donnée, on pourra alors envisager d'utiliser ces méthodes pour cartographier le «changement» des variables de l'enquête en fonction du changement de cette variable particulière. Il conviendra de noter à ce propos que de telles «cartes» constitueront des estimations non paramétriques des moyennes conditionnelles des variables de l'enquête, compte tenu de cette variable d'«ordonnement» ou de «lissage». L'analyste devra toutefois tenir compte du grave problème de la dimensionnalité: la taille de l'échantillon effectif diminue radicalement avec chaque nouvelle variable de lissage utilisée avec ces techniques non paramétriques.

Finalement, pour la cartographie des données d'enquête, nous avons utilisé des méthodes d'estimation fondées sur le noyau. Toutefois, on pourrait également recourir au lissage «spline», ou même à des méthodes paramétriques. À notre avis, le choix de la technique de lissage est quelque peu subjectif et dépend du but visé, puisqu'il n'y a pas de raisons objectives définitives pour préférer l'une ou l'autre méthode.

REMERCIEMENTS

Les auteurs tiennent à remercier les examinateurs pour leurs utiles commentaires qui ont grandement amélioré la présentation du présent article.

ANNEXE

Au cours des dernières années, on a établi un certain nombre de propriétés d'optimalité pour les poids du noyau localement linéaires (voir par exemple Wand et Jones (1995) et les articles auxquels ils font références). Nous avons par conséquent comparé les séquences de poids de Nadaraya-Watson (NW) et les séquences localement linéaires (LL) en utilisant des largeurs de bande fixes (LBF) et les largeurs de bande du k-ième plus proche voisin (PPV) avec chaque séquence de poids. Pour chacune de ces combinaisons, nous avons choisi la largeur de bande en utilisant la contre-validation des moindres carrés (CV) ou une méthode spéciale (décrite dans le dernier paragraphe de la présente section) destinée à atténuer l'aspect tacheté des cartes (AATC).

Deux critères nous ont servis à évaluer la performance de chaque méthode. Le premier, l'erreur quadratique moyenne (EQM), est le critère statistique évident pour l'évaluation d'un estimateur biaisé. Le deuxième critère est plus particulier au ABARE. À mesure qu'on produit les estimations en tableaux (par État) et en cartes, l'impression que nous laissons la carte en ce qui concerne la moyenne pour l'État devrait être proche de la valeur indiquée au tableau. Nous avons donc utilisé une somme pondérée du carré des différences observées entre les moyennes pour l'État des données brutes et des données lissées (EB<sup>2</sup>). Ce paramètre a également été calculé au niveau régional (RB<sup>2</sup>, chaque État contenant de une à neuf régions).

Les données ont été générées pour chaque emplacement en utilisant trois fonctions au degré de lissage différent (mesurées par  $\int m''$ ) et des erreurs de mélange normal. Par exemple,

$$m_1(z) = 6.25 \times 10^4 \times \cos \left( \frac{z_1 - 132.5}{2.25} \right) \cos \left( \frac{z_2 + 27.5}{1.75} \right),$$

où  $z_1$  et  $z_2$  représentent la longitude et la latitude du point  $z$ . Les fonctions  $m_i(z)$  ont été mises à l'échelle de manière à présenter la même étendue que les valeurs lissées d'une variable d'enquête clé, et les erreurs ont été mises à l'échelle pour présenter la même étendue que les résiduelles de la même variable après le lissage. Les emplacements caractérisés par des résiduelles importantes ont donné des valeurs élevées de la variance, et celles dont les résiduelles étaient faibles ont donné de petites valeurs de la variance. Les résultats de la simulation fondés sur la fonction de lissage sont présentés au tableau 5.

Tableau 5

Comparaison des séquences de poids localement linéaires (LL) et de Nadaraya-Watson (NW), utilisant des largeurs de bande fixes (LBF) et du k-ième plus proche voisin (PPV) choisies par contrevalidation des moindres carrés (CV) et à l'aide du critère décrit ci-après (AATC). Les résultats ont été obtenus à partir de 400 échantillons indépendants avec la fonction moyenne et des erreurs de mélange normales. Les valeurs de l'erreur quadratique moyenne ont été calculées à l'aide de la moyenne de la population finie de  $(y - m(z))^2$

		EQM $\times 10^{-7}$		RB <sup>2</sup> $\times 10^{-7}$		EB <sup>2</sup> $\times 10^{-7}$	
		CV	AATC	CV	AATC	CV	AATC
LL	LBF	39.64	93.93	4.44	1.67	1.33	0.39
	PPV	20.50	22.83	2.22	1.35	0.37	0.14
NW	LBF	41.91	52.78	3.29	1.77	0.34	0.17
	PPV	21.77	22.22	3.03	2.33	0.62	0.41

L'utilisation de l'erreur quadratique moyenne en guise de critère d'évaluation de la méthodologie n'a pas donné de résultats cohérents pour les trois fonctions  $m_i(z)$ . Toutefois, lorsque les valeurs RB<sup>2</sup> ou EB<sup>2</sup> ont été utilisées pour la mesure de la performance, l'estimateur LL avec une largeur de bande du k-ième plus proche voisin, sélectionné à l'aide de



Toutefois, elles n'offrent pas la même interprétation de l'échantillonnage répétée que les intervalles de confiance et devraient donc servir d'indicateurs plutôt que de mesures de l'incertitude associées à une carte particulière.

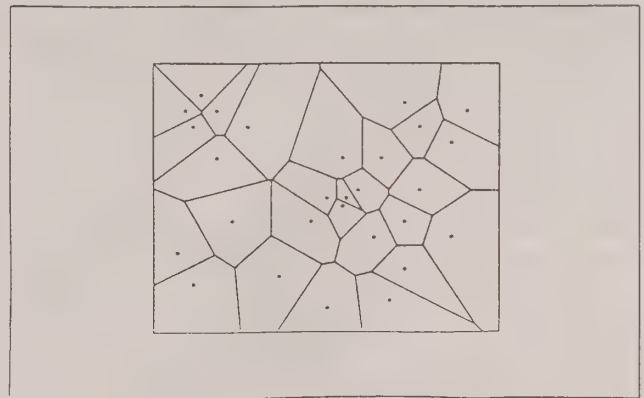
Pour des raisons de confidentialité, il convient de prendre

garde, lors de la préparation des cartes de données lissées aux fins de publication, de ne pas révéler l'emplacement des exploitations étudiées. Il convient en outre d'assurer une qualité des résultats qui soit compatible avec les systèmes d'éditique. Nous avons mis au point deux méthodes permettant de générer les cartes finales répondant à ces exigences en utilisant le SIG ARC/INFO.

Dans la première méthode, un polygone de Thiessen est

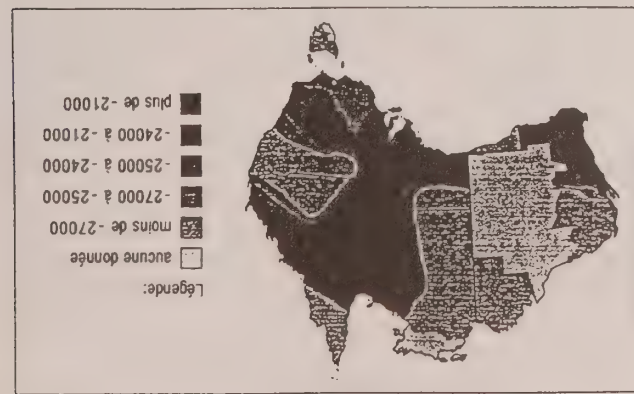
construit autour de chaque exploitation. Ce polygone délimite l'espace qui se trouve plus proche de la ferme en question que de toutes les autres. La ferme n'est pas au centre de ce polygone, et le périmètre du polygone ne suit pas les limites de l'exploitation, ce qui permet de dissimuler l'emplacement des fermes étudiées comme l'illustre la figure 7. Chaque polygone est coloré selon la valeur lissée de  $X$  à l'emplacement de la ferme dans ce polygone. On utilise habituellement dix couleurs pour chaque carte, et les déciles estimés de la population des données lissées servent de limites à la zone colorée. Les cartes présentées dans le présent article sont des analogues en noir et blanc de ces cartes en couleur.

**Figure 7.** Polygones de Thiessen construits autour d'exploitations choisies pour l'enquête du ABARE. L'emplacement de la ferme est indiqué par un petit carré à l'intérieur de chacun des polygones



Dans la deuxième méthode, les valeurs lissées d'une grille rectangulaire dense remplacent les valeurs lissées à l'emplacement de la ferme et on procède à une interpolation minime supplémentaire des données dans le système ARC/INFO. On établit une surface continue tridimensionnelle passant par les valeurs lissées des points de la grille, en deux étapes. Dans la première approximation, on construit une surface en facettes triangulaires à l'aide de la triangulation de Delauney pour ajuster ensuite un polynôme bivariable du cinquième degré à l'intérieur de chacun de ces triangles à l'aide de l'algorithme de Akima (Akima 1978). On détermine ensuite les contours de la surface continue résultante en utilisant les déciles estimés de la population. Cette méthode est illustrée à la figure 8.

**Figure 8.** Carte d'isolignes des profits des exploitations agricoles en 1991-1992, pour l'ensemble des grandes exploitations (\$)



Avec cette deuxième méthode de présentation, les emplacements des fermes visées par l'enquête ne sont pas pris en compte, ce qui permet de cacher complètement la provenance des données. On obtient en outre des contours plus lissés et les résultats ne sont pas aussi morcelés que sur les cartes en polygones. En outre, le personnel graphique du ABARE préfère cette méthode puisqu'elle réduit le nombre des sections à colorier séparément et qu'elle présente des exigences de stockage moindres, ce qui facilite la manipulation des cartes avec les logiciels d'édition. Elle a cependant pour inconvénient d'allonger le temps de traitement à l'étape du SIG. Puisque les méthodes décrites plus haut font l'interpolation des données sur l'ensemble du territoire australien, y compris les zones où on ne pratique pas l'agriculture, l'étape finale de la production des cartes avec le système ARC/INFO consiste à masquer les zones où le nombre d'exploitations pratiquant le type de production représentée par la carte est réduit ou nul. Comme le montre la figure 9, des régions différentes sont ainsi masquées selon le type de production visé.

**Figure 9.** Carte en polygones illustrant la variation probable de la production de laine, de 1991-1992 à 1992-1993, pour les exploitations possédant 100 moutons ou plus en 1991-1992 (kg)

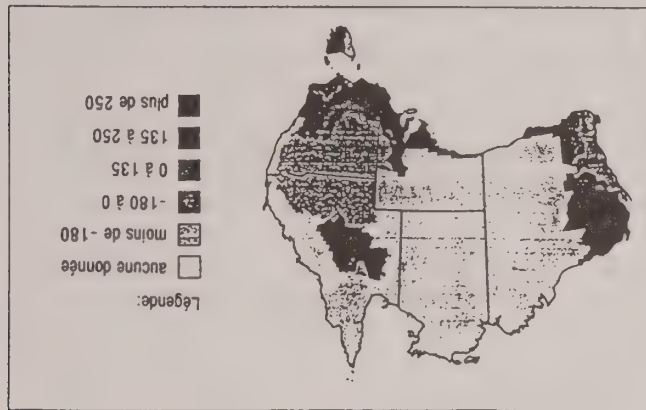




Tableau 4  
Biais et variance asymptotiques des estimateurs de

Nadaraya-Watson:

$$c_K = \int K^2(u)du, d_K = \int u^2 K(u)du$$

Largueur de bande fixe k-ième plus proche voisin

Biais

$$h^2 \frac{m''_f + 2m''_f(x)}{2f(x)} d_K \left( \frac{n}{k} \right)^2 \frac{(m''_f + 2m''_f(x))}{8f^3(x)} d_K$$

Variance

$$\frac{\sigma^2(x)}{n h f(x)} c_K \frac{k}{2\sigma^2(x)} c_K$$

Il apparaît clairement que le biais de la fonction de régression estimée peut être réduit en utilisant une largeur de bande plus étroite  $h$  (nombre de plus proches voisins  $k$ ), mais ceci conduit à une estimation bruitée  $\hat{m}$  dont les détails locaux masquent les caractéristiques globales de la courbe ( $\hat{m}$  est assorti d'une grande variance). Si  $h(k)$  est grand,  $\hat{m}$  sera plus lisse, mais les caractéristiques globales seront amorties ( $\hat{m}$  est assorti d'un biais élevé et d'une faible variance). Ainsi, le biais ne peut être réduit qu'aux dépens de la variance, et vice versa, la valeur  $h$  de la largeur de bande déterminant le rapport du biais (au carré) sur la variance.

En réalité, le plan d'échantillonnage et la distribution spatiale d'une variable d'enquête  $X$  ne seront pas indépendants, et les moyennes locales simples pour  $X$  dérivées des données de l'échantillon ne permettront pas d'obtenir de bonnes estimations des moyennes de la population locale de cette variable. Pour contourner ce problème, les poids du noyau sont multipliés par les poids de l'enquête pour donner les poids du lissage final utilisés dans le calcul de la moyenne locale. Cette opération équivaut à calculer la valeur estimée de la moyenne de la population locale  $m(z)$  de  $X$ , en tenant pour acquis qu'elle est localement linéaire pour les mêmes variables repères que celles utilisées dans la modélisation de la moyenne globale de la population de  $X$ .

Il a déjà été question, dans la documentation spécialisée, d'une vaste gamme de méthodes de lissage du noyau. Outre les diverses séquences de poids de lissage  $\{W_i\}$ , il existe différents types de largeurs de bandes et de nombreuses méthodes de sélection automatique de la largeur de bande. Nous avons donc procédé à une simulation destinée à déterminer la méthodologie fondée sur le noyau la plus appropriée pour l'établissement des cartes du ABARE. Cette étude est décrite en annexe.

L'incertitude qui entoure l'estimation de la moyenne spatiale dérivée par un lissage spatial fondé sur le noyau peut être illustrée par la cartographie de la variabilité locale de la variable d'intérêt. Les zones de variabilité locale élevée correspondent aux régions où la carte de la fonction moyenne est moins précise, et vice versa pour les zones de variabilité locale faible.

La méthode habituelle de détermination des régions de confiance pour une estimation de la courbe du noyau a reçu le nom de «méthode bootstrap» (voir Härdle 1990; Hall 1992 et les ouvrages cités en référence dans ces articles). Toutefois, pour des raisons d'efficacité de calcul, nous utilisons les

«expectiles» (Newey et Powell 1987) de la distribution spatiale de  $X$  pour décrire cette variabilité locale. Un expectile présente un rapport à la moyenne équivalant au rapport du quartile à la médiane. En particulier, la différence entre le 75<sup>e</sup> et le 25<sup>e</sup> expectiles d'une distribution constitue une mesure de l'étalement de la distribution comparable à l'interquartile. Le programme de lissage contient un module pour la régression  $M$ -quantile non paramétrique (Breckling et Chambers 1988) qu'on utilise pour ajuster la surface lissée aux expectiles de la distribution  $X$  à un endroit donné. La différence entre les surfaces lissées du 75<sup>e</sup> et du 25<sup>e</sup> expectiles (l'expectile lissé analogue à l'étendue de l'interquartile) est alors cartographiée pour montrer les zones de haute et de basse variabilité des données.

On constate sans surprise que cette étendue lissée de l'interexpectile tend à être plus élevée dans les régions où les exploitations sont éloignées les unes des autres et où la variabilité de  $X$  est donc la plus élevée. Nous présentons à la figure 6 la carte de l'étendue de l'interexpectile correspondant à la figure 5. Notez que ces cartes de l'étendue de l'interexpectile lissé fournissent des informations semblables aux bandes de confiance à n'importe quel point particulier de la carte.

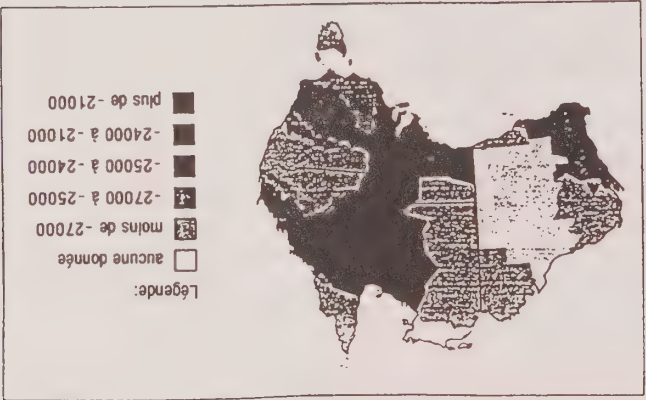


Figure 5. Carte en polygones du profit des exploitations agricoles en 1991-1992, sur l'ensemble des grandes exploitations (\$)

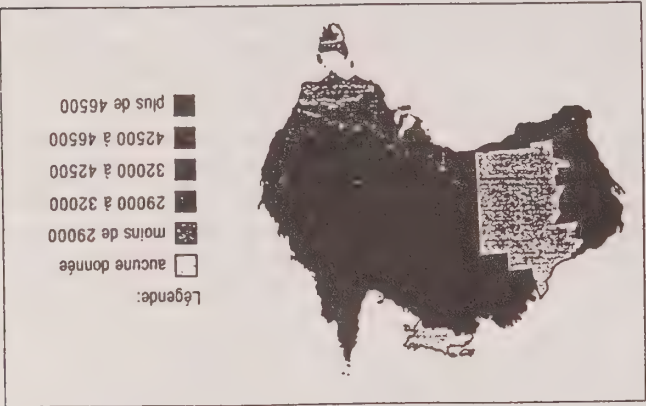


Figure 6. Carte en polygones de l'étendue de l'interexpectile du profit des exploitations agricoles en 1991-1992, sur l'ensemble des grandes exploitations (\$)

Tableau 3  
Estimations (valeurs estimées de l'écart-type entre parenthèses) de la valeur moyenne de  $X$  = collis totaux au comptant dans les sous-régions SR-1 à SR-7, constituant la Région B (taille de l'échantillon  $n$  = 85 exploitations), obtenues avec des poids d'écrétage standards (2.3) et des poids d'écrétage à lissage spatial (2.6)

	Poids d'écrétage à lissage spatial		Poids d'écrétage à lissage spatial					
	Poids standards	$d = 0.05$	$d = 0.05$	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 0.9$	$\alpha = 0.9$	$\alpha = 0.1$
SR-1	183,194	183,262	183,528	186,151	184,287	195,138	257,652	(64,851)
	(64,851)	(64,325)	(64,051)	(64,967)	(64,132)	(69,859)	(59,518)	
SR-2	261,952	261,487	261,119	261,182	261,938	276,912	331,805	(70,989)
	(70,989)	(70,601)	(70,502)	(73,131)	(70,723)	(79,751)	(67,356)	
SR-3	113,499	113,441	113,742	116,847	114,631	125,525	157,007	(30,304)
	(30,304)	(30,289)	(30,255)	(30,731)	(30,377)	(31,507)	(32,500)	
SR-4	242,220	242,182	242,208	242,221	242,163	242,439	250,871	(26,160)
	(26,160)	(25,671)	(26,159)	(26,160)	(26,154)	(24,244)	(24,836)	
SR-5	134,524	134,970	135,700	139,122	134,734	131,448	148,629	(32,420)
	(32,420)	(32,528)	(32,432)	(30,607)	(32,202)	(27,942)		
SR-6	176,540	176,977	175,708	163,241	172,076	148,434	171,856	(60,377)
	(60,377)	(60,703)	(59,214)	(46,361)	(55,925)	(36,218)		
SR-7	205,287	205,644	205,433	202,039	204,519	194,998	219,959	(44,137)
	(44,137)	(44,008)	(43,963)	(44,044)	(43,972)	(45,434)	(51,690)	
Région B	176,283	176,342	176,397	176,822	176,294	179,998	216,445	(19,039)
	(19,039)	(18,869)	(18,874)	(18,213)	(18,511)	(18,540)	(17,099)	

#### 4. ESTIMATION ET CARTOGRAPHIE DES MOYENNES LOCALES

Une carte de données d'enquête est une surface bidimensionnelle qui permet d'estimer la fonction moyenne spatiale de la variable d'enquête dans la population. En pratique, on obtient une telle carte en appliquant des techniques de régression non paramétriques aux données d'enregistrement unitaires pondérées obtenues dans le cadre de l'enquête.

Le ABARE utilise la régression de noyau (une technique non paramétrique) pour produire des cartes qui montrent la variation spatiale des surfaces de la fonction moyenne spatiale estimée de variables d'enquête clés. Ces surfaces sont obtenues en remplaçant les valeurs d'échantillons observées de ces variables par des moyennes localement pondérées. En outre, pour chaque carte de moyenne locale, on produit une carte correspondante qui donne une estimation de la variabilité locale de la variable d'intérêt. Nous présentons ci-après un bref aperçu de la technique: pour des raisons de simplicité, nous traiterons uniquement du cas univarié. Voir Ruppert et Wand (1994), Wand et Jones (1995, p.140) et les articles auxquels ils font référence pour en savoir plus sur le cas multivarié.

Nous présentons que la population finie générée est un échantillon iid  $\{(Z_i^j, Y_i^j), i = 1, \dots, N\}$  tiré d'une super-population où  $Y_i^j$  est la valeur d'une variable de réponse  $Y$  observée au lieu  $Z_i^j$ . Nous supposons que les observations obéissent au modèle

$$Y_i^j = m(Z_i^j) + \epsilon_i^j, \quad i = 1, \dots, N$$

où  $m(z) = E(Y|Z = z)$  correspond à la moyenne conditionnelle de  $Y_i^j$  étant donné  $Z_i^j$ , et  $\epsilon_i^j$  désigne les variables aléatoires indépendantes ayant une moyenne zéro et une variance  $\sigma^2(z)$ . Supposons que les termes d'erreurs  $\epsilon_i^j$  sont indépendants du processus de sélection de l'échantillon, de sorte que les valeurs de l'échantillon  $\{(Z_i^j, Y_i^j), i = 1, \dots, n\}$  obéissent au même modèle, et désignons par  $f$  la densité de  $Z_1^j, \dots, Z_n^j$ . Le choix naturel de la moyenne locale d'un point  $z$  quelconque devient alors la moyenne des valeurs de la variable réponse pour les observations dont l'emplacement est proche de  $z$ , puisque les observations de points éloignés ont tendance à montrer des valeurs moyennes très différentes. La moyenne locale est donc définie comme une moyenne pondérée

$$\hat{m}(z) = n^{-1} \sum_{i=1}^n W_i^j(z) Y_i^j$$

où les poids  $\{W_i^j(z)\}$  dépendent des emplacements  $\{Z_i^j\}$  des observations de l'échantillon, et où  $\hat{m}(z)$  est une estimation de  $m(z)$ .

Les poids sont élaborés à l'aide d'une fonction  $K$  connue sous le nom de noyau, continue, bornée, symétrique et dont l'intégrale est 1. Diverses séquences de poids ont déjà été proposées: les poids traditionnels de Nadaraya-Watson (Nadaraya 1964 et Watson 1964) sont

$$W_i^j(z) = h^{-1} K\{(z - Z_i^j)/h\} \left/ \sum_{j=1}^f \left[ (nh)^{-1} K\{(z - Z_j^j)/h\} \right] \right.$$

où  $h$  est un facteur scalaire appelé largeur de bande. La fonction  $K$  donne à une observation proche de  $z$  une influence relativement plus grande sur la moyenne locale à cet emplacement qu'elle ne le fait pour une observation plus éloignée de  $z$ . Lorsque le nombre d'observations est limité, une fenêtre à largeur de bande fixe pourrait contenir un nombre limité de points et l'estimateur correspondant risquerait donc d'avoir une variance très élevée. On peut éviter ce problème en utilisant la méthode du  $k$ -ième plus proche voisin dans laquelle la largeur de bande diffèrentielle est utilisée pour chaque estimation du point  $z$ . La largeur de bande à  $z$  est la distance au  $k$ -ième plus proche voisin de  $z$ , de sorte qu'il y aura toujours exactement  $k$  points dans la fenêtre. Désignons par  $h_k$  la distance qui sépare  $z$  de son  $k$ -ième plus proche voisin. Les poids  $k$ -ième plus proche voisin de Nadaraya-Watson sont

$$W_{h_k}^j(z) = h_k^{-1} K\{(z - Z_i^j)/h_k\} \left/ \sum_{j=1}^f \left[ (nh_k)^{-1} K\{(z - Z_j^j)/h_k\} \right] \right.$$

Nous présentons au tableau 4 les propriétés de l'erreur quadratique moyenne (EQM) asymptotique des estimateurs habituels (largeur de bande fixe) et des estimateurs  $k$ -ième plus proche voisin, tels qu'établis par Härdle (1990, p. 46).



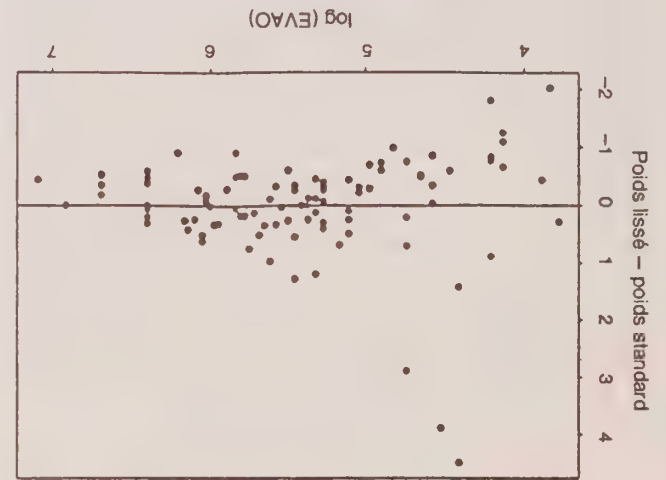
à un changement la hausse des estimations des enquêtes pour ces régions avec l'introduction de poids d'échantillons à fort lissage spatial. Compte tenu du biais positif accru indiqué au tableau 1, cette augmentation devrait être essentiellement due à l'introduction d'un biais positif dans ces estimations.

Cette augmentation du biais est-elle compensée par un écart-type plus faible? Pour répondre à cette question, nous avons calculé les valeurs estimées de l'enquête et les écarts-types correspondant à une variable financière clé: le total des coûts au comptant. Ces estimations sont compilées aux tableaux 2 (région A) et 3 (région B). On fournit les estimations pour chacune des régions ainsi que pour les sous-régions comprises à l'intérieur de chacune d'elles et qui sont identifiées dans les tableaux par SR-1, l'indice *i* variant de 1 à 6 pour la région A et de 1 à 7 pour la région B.

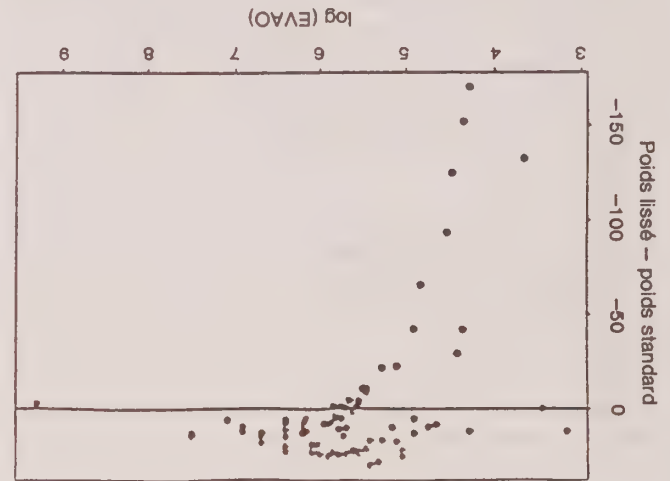
On observe qu'en général, la réponse à la question posée ci-dessus est affirmative. Les valeurs estimées de l'écart-type des estimations de l'enquête diminuent à mesure qu'augmente le degré de lissage spatial des poids (de gauche à droite, dans les tableaux). Toutefois, comme prévu, les estimations augmentent elles aussi en taille, montrant un biais positif de plus en plus important. Dans l'ensemble, l'avantage procuré par une baisse de l'écart-type semble compenser pour l'augmentation du biais, sauf dans les cas de lissage spatial plus important ( $\alpha = 0,1$ ,  $d = 0,005$ ). Dans ce dernier cas, l'augmentation du biais dépasse la réduction de l'écart-type. Le choix d'un  $\alpha = 0,1$  et d'un  $d = 0,05$  paraît être un compromis acceptable, permettant d'obtenir un équilibre raisonnable (mais non spectaculaire) entre le biais et la variance dans la région A, et peu de changements dans les estimations de la région B.

**Tableau 2**  
Estimations (valeurs estimées de l'écart-type entre parenthèses) de la valeur moyenne de  $X$  = coûts totaux au comptant dans les sous-régions SR-1 à SR-6, constituant la Région A (taille de l'échantillon  $n = 101$  exploitations), obtenues avec des poids d'écrêtage standards (2,3) et des poids d'écrêtage à lissage spatial (2,6)

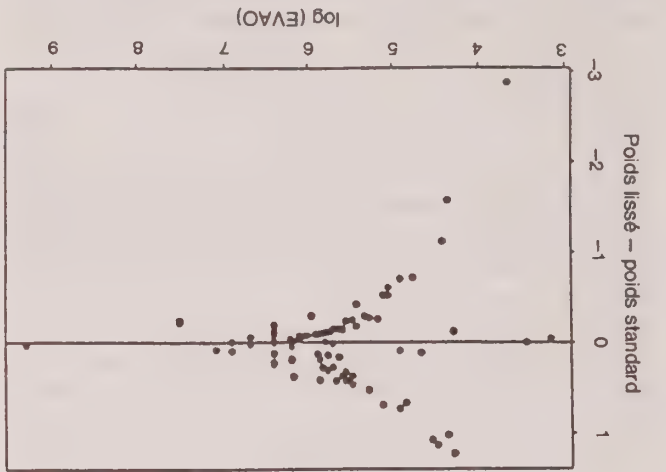
	Poids standards						Poids d'écrêtage à lissage spatial					
	$\alpha = 0,9$	$\alpha = 0,5$	$\alpha = 0,1$	$\alpha = 0,9$	$\alpha = 0,5$	$\alpha = 0,1$	$d = 0,05$	$d = 0,05$	$d = 0,05$	$d = 0,05$	$d = 0,05$	$d = 0,05$
SR-1	100,618	100,453	101,297	107,263	102,059	112,635	135,419	(24,551)	(23,511)	(23,906)	(20,487)	(18,011)
SR-2	115,320	115,417	116,002	120,362	116,917	126,165	153,707	(26,661)	(26,448)	(25,637)	(26,423)	(27,975)
SR-3	167,524	167,453	167,486	168,257	167,709	170,781	187,683	(28,479)	(28,467)	(28,473)	(28,175)	(24,211)
SR-4	182,940	180,317	177,838	163,556	176,257	174,077	192,296	(106,471)	(105,485)	(101,012)	(74,418)	(43,651)
SR-5	132,050	132,083	132,389	134,786	132,490	136,369	151,046	(25,089)	(25,096)	(25,154)	(25,475)	(23,110)
SR-6	132,493	132,184	132,204	141,623	133,763	147,652	192,781	(44,385)	(44,546)	(44,757)	(45,078)	(53,105)
Région A	134,114	133,807	134,141	137,080	134,506	142,040	166,432	(15,691)	(15,655)	(15,426)	(13,845)	(12,815)
								(13,494)	(13,494)	(13,494)	(13,494)	(12,815)



**Figure 2.** Différence entre le poids lissé avec  $\alpha = 0,9$  et  $d = 0,05$  et le poids d'écrêtage standard, Région A



**Figure 3.** Différence entre le poids lissé avec  $\alpha = 0,1$  et  $d = 0,005$  et le poids d'écrêtage standard, Région B



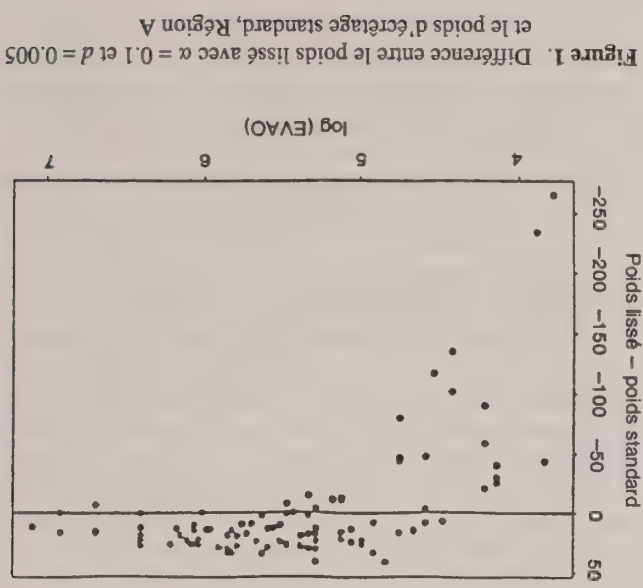
**Figure 4.** Différence entre le poids lissé avec  $\alpha = 0,9$  et  $d = 0,05$  et le poids d'écrêtage standard, Région B

exploitations les plus grandes de l'échantillon, tout en diminuant radicalement les poids d'un petit nombre d'exploitations plus petites. Le lissage spatial faible (figures 2 et 4) influe beaucoup moins sur les poids, et le rapport entre la taille de l'exploitation et la direction du changement des poids n'est pas très évident. En conséquence, on devrait s'attendre

**Tableau 1**  
Valeurs (en pourcentages relatifs) des biais associés à l'estimation des variables repères correspondant aux principaux produits agricoles de la région A (taille de l'échantillon  $n = 101$  exploitations) et de la région B (taille de l'échantillon  $n = 85$  exploitations) obtenues avec des poids d'écrétage standards (2.3) et des poids d'écrétage à lissage spatial (2.6)

Région A			
Poids d'écrétage standards	Ble	Moutons	Riz
$d = 0.05$	-0.50	4.6	11.9
$\alpha = 0.9$	-0.46	4.7	12.4
$\alpha = 0.5$	0.07	6.2	17.4
$\alpha = 0.1$	-0.40	4.9	12.7
$d = 0.005$	-0.40	0.80	28.0
$\alpha = 0.9$	9.20	25.0	60.0
$\alpha = 0.5$			
$\alpha = 0.1$			
Région B			
Poids d'écrétage standards	Ble	Moutons	Légumineuses
$d = 0.05$	0.42	-1.16	1.37
$\alpha = 0.9$	0.44	-1.14	1.40
$\alpha = 0.5$	0.69	-1.25	2.53
$\alpha = 0.1$	0.50	-1.20	1.68
$d = 0.005$	1.51	19.61	45.46
$\alpha = 0.5$			
$\alpha = 0.1$			

Région B			
Poids d'écrétage standards	Ble	Moutons	Légumineuses
$d = 0.05$	0.42	-1.16	1.37
$\alpha = 0.9$	0.44	-1.14	1.40
$\alpha = 0.5$	0.69	-1.25	2.53
$\alpha = 0.1$	0.50	-1.20	1.68
$d = 0.005$	1.51	19.61	45.46
$\alpha = 0.5$			
$\alpha = 0.1$			



**Figure 1.** Différence entre le poids lissé avec  $\alpha = 0.1$  et  $d = 0.005$  et le poids d'écrétage standard, Région A

où  $\|z_i - z_j\|$  représente la distance qui sépare l'exploitation  $i$  de l'exploitation  $j$ , et  $d$  est une constante de contrôle du rayon du cercle entourant la  $i$ -ième exploitation, à l'intérieur duquel on a procédé au lissage spatial. Plus la valeur de  $d$  est petite, plus grand sera le rayon du lissage spatial. A l'heure actuelle, la constante de multiplication  $\phi$  correspond au rapport des déterminants des matrices  $K$  et  $\omega$ , élevées à la puissance  $n^{-2}$ . Nous procédons, à la section suivante, à une évaluation empirique de cette méthode.

### 3. APPLICATION DE LA PONDERATION D'ÉCHANTILLON À LISSAGE SPATIAL

Les résultats initiaux d'une évaluation de la première méthode de pondération par écrétage à lissage spatial décrits dans la section précédente sont présentés aux tableaux 1 à 3.

Ils correspondent à deux régions agricoles distinctes. La première, appelée A, se trouve en Nouvelle-Galles du Sud. En termes d'espace, cette région est relativement homogène, étant située dans la portion sud-ouest de l'Etat. La production de blé et de riz et la production de laine et d'agneau y sont les principales activités agricoles. La deuxième région, appelée B, se trouve en Australie-Occidentale. C'est une région plus hétérogène, comportant des exploitations de productions végétales et de production de laine, dans le centre-ouest de l'Etat, et des exploitations beaucoup plus vastes d'élevage du bétail et de cultures installées sur des terres agricoles marginales dans le sud-est. On y produit principalement du blé et des légumineuses, ainsi que de la laine.

Nous avons utilisé six variations des poids d'écrétage à lissage spatial (2.6) avec une valeur de  $K$  donnée par (2.7), définie par les valeurs de  $d = 0.05$  (effets spatiaux faibles) et  $d = 0.005$  (effets spatiaux forts), et avec des valeurs de  $\alpha = 0.9$  (accent sur les poids d'écrétage standards) et  $\alpha = 0.5$  (importance égale accordée aux poids d'écrétage standards et aux poids à lissage spatial) et  $\alpha = 0.1$  (accent sur les poids à lissage spatial).

Le tableau 1 présente les valeurs relatives du biais associées à l'estimation des valeurs repères totales liées au produit principal pour chaque région, en vertu de ces différents systèmes de pondération, ainsi que les valeurs correspondantes du biais associées aux poids d'écrétage standards. L'augmentation de la valeur de ces biais à mesure qu'on augmente le lissage spatial des poids est évidente. Comme il existe une corrélation positive entre ces repères de production et la plupart des variables économiques mesurées dans le cadre de l'enquête, on peut s'attendre à ce que ces biais repères se traduisent par une augmentation correspondante des biais des estimations de l'enquête fondés sur ces poids. Les figures 1 à 4 illustrent les différences qui existent entre les poids lissés et les poids d'écrétage standards pour les deux combinaisons «extrêmes» de  $\alpha$  et  $d$  dans les deux régions, qui changent à mesure que change la taille des exploitations échantillonnées (mesure correspondant au logarithme de la valeur estimée des opérations agricoles, ou  $\log(EVAO)$ ). Soulignons qu'un lissage spatial relativement fort (figures 1 et 3) a pour effet d'accroître les poids de la plupart des



Les travaux récents du ABARE portant sur l'estimation pour les petites régions ont surtout cherché à modifier cette méthode de pondération par échantillon afin de créer des poids d'échantillons dont la variabilité spatiale serait moindre. Nous y parvenons en modifiant le critère  $\bar{Q}$  de l'erreur quadratique moyenne dans (2.2) afin d'inclure une contrainte de la variabilité spatiale, tout en continuant de considérer les éléments de la variable  $X$  comme des variables indépendantes. Désignons par  $K$  une matrice  $n \times n$  reflétant la distance Euclidienne entre les exploitations agricoles échantillonnées, de sorte que  $K$  soit symétrique et non négative,  $K_{ii} = 1$  pour toutes les valeurs de  $i$  et  $K_{ij} \downarrow 0$  à mesure que la distance qui sépare l'exploitation  $i$  de l'exploitation  $j$  augmente. Supposons que  $u = w - 1$ . La valeur de  $u$  que nous choisirons devra faire en sorte que lorsque  $K_{ij}$  est grand, la différence entre  $u_i$  et  $u_j$  est petite. Autrement dit, nous chercherons à minimiser une quantité de la forme

$$\sum_{i \neq j} K_{ij} (u_i - u_j)^2 = 2(n^{(2)})^T K 1 - 2u^T K u \quad (2.4)$$

où  $(n^{(2)})_i = (u_i)^2$ . Une modification appropriée du critère de l'erreur quadratique moyenne (2.2) conduit à la minimisation de

$$\bar{Q}^* = \lambda^{-1} B^T C B + u^T \omega u + (n^{(2)})^T K 1 - u^T K u.$$

La minimisation par rapport à  $u$  conduit à

$$u = \eta^{-1} x (\lambda C^{-1} + x^T \eta^{-1} x)^{-1} (T - x^T 1)$$

à condition qu'il existe une valeur  $\eta^{-1}$ , où

$$\eta = \text{diag}(K 1) - K + \omega. \quad (2.5)$$

Il apparaît ainsi clairement que

$$w = 1 + \eta^{-1} x (\lambda C^{-1} + x^T \eta^{-1} x)^{-1} (T - x^T 1). \quad (2.6)$$

On constate que le critère modifié de l'erreur quadratique moyenne  $\bar{Q}^*$  pondère à la fois le critère de lissage spatial donné dans (2.4) et le terme correspondant à la variance de l'erreur de prédiction des estimations de l'échantillon,  $u^T \omega u$ . Comme l'échelle de  $K$  a été spécifiée arbitrairement, la pondération comparative des deux critères doit être modifiée en multipliant la matrice spatiale  $\{\text{diag}(K 1) - K\}$  par un facteur  $\phi$ , afin de la rendre comparable par la taille à la matrice d'hétéroscédasticité  $\omega$ , et en ajoutant un paramètre  $\alpha$ ,  $0 \leq \alpha \leq 1$ , à l'expression de  $\eta$  dans l'équation (2.5) de sorte que

$$\eta = (1 - \alpha) \phi \{\text{diag}(K 1) - K\} + \alpha \omega.$$

On peut dériver ces poids d'échantillons à lissage spatial selon une autre méthode qui donne un meilleur aperçu de la façon dont ils devraient être interprétés. Cette méthode découle du fait que la matrice

$$\eta = \begin{bmatrix} \sigma_1^2 + \sum_{m \neq 1} K_{1m} & -K_{12} & \dots & -K_{1n} \\ -K_{21} & \sigma_2^2 + \sum_{m \neq 2} K_{2m} & \dots & -K_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ -K_{n1} & -K_{n2} & \dots & \sigma_n^2 + \sum_{m \neq n} K_{nm} \end{bmatrix}$$

peut être assimilée à l'équation  $\eta = S R S$ , où  $S$  est une matrice diagonale avec  $S_{ii} = (\sigma_i^2 + \sum_{m \neq i} K_{im})^{1/2}$ , et  $R$  est une matrice de corrélation avec

$$R_{ij} = \begin{cases} 1 & \text{si } i = j \\ -K_{ij} \left\{ \left( \sigma_i^2 + \sum_{m \neq i} K_{im} \right) \left( \sigma_j^2 + \sum_{m \neq j} K_{jm} \right) \right\}^{-1/2} & \text{si } i \neq j. \end{cases}$$

Ainsi, les poids d'échantillons à lissage spatial peuvent être dérivés comme des poids de régression de type à écrêtage en présumant que la variable  $X$  obéit à un modèle linéaire de la forme (2.1), avec une valeur de  $V$  redéfinie pour satisfaire

$$E(V) = 0, \text{ var}(X_i) = \sigma_i^2 + \sum_{m \neq i} K_{im}, \text{ et } \text{cov}(X_i, X_j) = -K_{ij}$$

pour  $i \neq j$ . La méthode habituelle de pondération par écrêtage mène ainsi directement à (2.6), avec une valeur de  $\eta$  définie par (2.5). À noter qu'en vertu de ce modèle, on obtient une corrélation négative entre les exploitations agricoles voisines.

Cette deuxième méthode de dérivation montre clairement que le recours au lissage spatial pour les poids de l'enquête s'accorde mal avec les notions standards d'efficacité statistique, dans la mesure où on s'intéresse à l'estimation au niveau de l'agrégat. Comme la corrélation spatiale entre les exploitations voisines sera typiquement positive, l'estimation efficace au niveau de l'agrégat comportera nécessairement une pondération fondée sur (2.3) où  $\omega$  sera remplacé par une matrice non diagonale variance/covariance reflétant cette corrélation spatiale positive. Ces poids ne sont pas ceux que l'on obtient lorsqu'on impose une contrainte de similarité spatiale. En conséquence, il faudra s'attendre à ce que ces poids «efficaces» pour les grandes régions» tendent à être plus différents pour les exploitations voisines qu'ils ne le seront pour les exploitations éloignées l'une de l'autre. Autrement dit, il y a un prix à payer pour la pondération: si nous avons besoin d'estimations au niveau de l'agrégat moins variables, ceci aura tendance à conduire à des estimations plus variables pour les petites régions. Inversement, si on choisit l'équation (2.6) aux fins de la pondération à cause de ses propriétés souhaitables pour les petites régions, il faudra s'attendre à ce que les estimations au niveau de l'agrégat obtenues en faisant la somme de ces estimations pour les petites régions soient moins efficaces.

Les poids d'échantillons à lissage spatial (2.6) ont été appliqués à l'aide de l'équation

$$K_{ij} = \exp(-d \|z_i - z_j\|), \quad (2.7)$$

généralement des poids extrêmement variables et souvent négatifs.

Comme l'expliquent Bardsley et Chambers (1984), il convient d'éviter les poids négatifs dans une enquête multivariante comme celle du AAGIS. De tels poids peuvent notamment conduire à des estimations négatives de quantités intrinsèquement positives. Ce problème a très souvent été évoqué dans la documentation spécialisée (voir par exemple Deville et Särndal 1992; Bankier, Rathwell et Majkowski 1992; et Fuller, Louglin et Baker 1994). La méthode utilisée par le ABARE pour garantir des poids d'échantillons positifs est fondée sur une modification du type par écrêtage des poids de la meilleure estimation linéaire non biaisée, tel que le suggèrent Bardsley et Chambers (1984).

Étant donné un échantillon de taille  $n$  provenant d'une région particulière, la méthode de pondération par écrêtage détermine la valeur du vecteur  $w$  du poids de l'échantillon en minimisant le critère de l'erreur quadratique moyenne:

$$\tilde{Q} = \lambda^{-1} B^T C B + (w - 1)^T w (w - 1). \quad (2.2)$$

Où  $B = T - x^T w$  est un  $p$ -vecteur des biais repères correspondant aux différences entre les totaux  $T$  de la population (comme) des variables repères  $p$  représentées par  $X$  et où les estimations d'enquête correspondent  $x^T w$  de ces totaux, représentées par  $C$ , constituent une matrice diagonale  $p \times p$  de «coûts» relatifs non négatifs associés à ces biais,  $w$  est une composante échantillon de  $\Omega$ ,  $x$  est une composante échantillon de  $X$ ,  $1$  est un  $n$ -vecteur de chiffres 1 et  $\lambda$  est une constante scalaire choisie par l'analyste d'enquête. La valeur de  $w$  qui minimise  $\tilde{Q}$  est

$$w = 1 + \omega^{-1} x (\lambda C^{-1} + x^T \omega^{-1} x)^{-1} (T - x^T 1). \quad (2.3)$$

La constante scalaire  $\lambda$  est le paramètre d'écrêtage associé à ces poids. À mesure que la valeur de  $\lambda$  augmente à partir de zéro, les poids de l'échantillon dans  $w$  s'éloignent de leurs meilleures valeurs linéaires non biaisées en vertu du modèle (2.1) (c.-à-d., de leurs valeurs lorsque  $\lambda = 0$ ) et voient leur variabilité diminuer graduellement. Autrement dit, à mesure que la valeur de  $\lambda$  augmente, les variances des estimations de l'enquête fondées sur ces poids diminuent. Cependant, à mesure que la valeur de  $\lambda$  augmente, ces estimations deviennent plus biaisées en vertu de (2.1), de sorte que les composantes de  $B$  s'éloignent de leurs valeurs zéro à  $\lambda = 0$  (où les poids d'échantillon définissent des estimations non biaisées en vertu de (2.1)). Ces composantes voient leur valeur augmenter graduellement (en termes absolus) à mesure que  $\lambda$  augmente. L'analyste d'enquête cherche une solution de compromis entre ces deux sources concurrentes d'«erreurs» en choisissant la plus petite valeur de  $\lambda$  qui permettra aux poids de l'échantillon dans  $w$  de se stabiliser à des valeurs strictement positives le plus proches possible de leurs meilleures valeurs linéaires non biaisées en vertu de (2.1). Ceci permet d'assurer que les composantes de  $B$  seront les plus petites possibles, sous réserve de cette exigence de stabilité. Le ABARE choisit cette valeur de  $\lambda$  de manière que les poids des échantillons soient au moins égaux à 1 unité.

La variation géographique d'une variable donnée et celle d'une autre variable. Finalement, les cartes en couleur ont une incidence extrêmement positive sur la présentation visuelle des données.

La demande accrue pour des informations spatiales nous a poussés à porter une attention particulière aux estimations portant sur des petites régions. Pour procéder à de telles estimations (qui découlent naturellement du lissage des données d'enquête aux fins de la présentation dans les cartes) on peut par exemple procéder au lissage spatial des poids d'échantillons. On réduit ainsi la variabilité des estimations portant sur de petites régions.

Dans la section 2, nous présentons une méthode qui permet d'intégrer l'emplacement géographique aux méthodes de pondération des enquêtes du ABARE afin de réduire la variabilité de nos estimations pour les petites régions. Dans la section 3, nous appliquons cette méthode aux estimations sous-régionales correspondant à deux régions agricoles. Dans la section 4, nous décrivons comment les techniques de régression du noyau peuvent servir à produire des cartes qui donneront une bonne indication de la variation géographique locale des variables d'enquête. Nous décrivons deux méthodes de cartographie des données lissées à l'aide du système d'information géographique ARC/INFO. Nous résumons finalement en annexe les résultats d'une étude de simulation comparant diverses méthodes de régression du noyau aux fins de l'utilisation dans les cartes du ABARE.

## 2. ESTIMATION POUR LES PETITES RÉGIONS AVEC LISSAGE SPATIAL DES POIDS D'ÉCHANTILLON

La méthode normalement utilisée par le ABARE pour le calcul des poids des échantillons est décrite par Bardsley et Chambers (1984). Elle est fondée sur l'hypothèse qui veut qu'à un certain degré d'aggrégation (p. ex., région agricole), la variable  $Y$  obéira à un modèle linéaire de la forme

$$Y = X\beta + V \quad (2.1)$$

où  $Y$  est le  $N$ -vecteur des valeurs de  $Y$  à ce niveau d'aggrégation,  $X$  est une matrice  $N \times p$  des valeurs d'un ensemble de  $p$  variables repères,  $\beta$  est un  $p$ -vecteur de coefficients de régression inconnu et  $V$  est un  $N$ -vecteur d'erreurs satisfaisant aux équations  $E(V) = 0$  et  $\text{var}(V) = \sigma^2 \Omega$ , où  $\sigma$  est un paramètre scalaire inconnu et  $\Omega$  est une matrice diagonale connue  $N \times N$  ayant pour éléments la mesure de l'importance de chaque exploitation, c'est-à-dire la EVAO définie dans la section précédente.

Puisque ce modèle est multivalent, le même groupe de variables repères étant utilisé pour chaque variable de l'enquête, la dimension colonne  $p$  de  $X$  est habituellement grande. Typiquement,  $X$  représentera entre 3 et 7 variables liées aux principaux produits agricoles des exploitations de la région à l'intérieur de la région. La meilleure estimation linéaire non biaisée du total de la population d'une variable d'enquête établie à partir d'un modèle à ce point hyperspécifié donnera



# Applications du lissage spatial aux données d'enquête

ANN COWLING, RAY CHAMBERS, RAY LINDSAY et BHAMATHY PARAMESWARAN<sup>1</sup>

## RÉSUMÉ

Dans le présent article, nous décrivons deux applications du lissage spatial en utilisant des données recueillies dans le cadre d'une enquête économique à grande échelle portant sur les exploitations agricoles australiennes: la première pour les petites régions et la seconde pour les grandes. Dans le premier cas, (petites régions), nous décrivons comment le lissage spatial des poids de l'échantillon peut permettre d'améliorer les estimations. Dans le second cas (grandes régions), nous proposons une méthode de lissage spatial et une méthode de cartographie des données lissées. La méthode standard de pondération utilisée pour l'enquête est une variante de la pondération à régression linéaire. Pour les petites régions, cette méthode est modifiée par l'introduction d'une contrainte sur la variabilité spatiale des poids. Les résultats d'une étude empirique à petite échelle laissent constater que cette méthode réduit comme prévu la variance des estimateurs des petites régions, mais au coût d'une augmentation de leur biais. Pour l'application aux grandes régions, nous décrivons la méthode de régression non paramétrique utilisée pour le lissage spatial des données d'enquête, ainsi que les techniques de cartographie de ces données lissées fondées sur un système d'information géographique (SIG). Nous présentons en outre les résultats d'une étude de simulation réalisée afin de déterminer la méthode et le degré de lissage les plus appropriés pour l'utilisation avec les cartes.

**MOTS CLÉS:** Estimations pour les petites régions; pondération d'enquête; estimation du noyau; cartographie des données d'enquête.

## 1. INTRODUCTION

Le Australian Bureau of Agricultural and Resource Economics (ABARE) est une organisation de recherches économiques appliquées rattachée au Department of Primary Industries and Energy. Le ABARE est notamment responsable des enquêtes annuelles conduites auprès d'industries agricoles australiennes choisies qui lui procurent une vaste gamme d'informations sur les caractéristiques économiques

et physiques des exploitations agricoles. La plus importante de ces enquêtes est le Australian Agricultural and Grazing Industries Survey (AAGIS), qui porte sur les exploitations agricoles dont la valeur estimée des opérations (EVAO) s'établit à \$22,500 ou plus pour la période visée par le dernier recensement agricole et qui font partie d'une des grandes industries de production agricole (broadacre industries): production céréalière, production de bovins de boucherie et production d'ovins et de laine. Au cours des deux dernières années, environ 1,650 de ses exploitations ont été incluses dans l'échantillon du AAGIS, lequel est stratifié par région géographique, par industrie et par EVAO. Les exploitations comprises sont inégalement réparties sur le territoire australien. Les renseignements normalement recueillis comprennent la latitude et la longitude des exploitations échantillonnées (correspondant à l'emplacement de l'entée principale). Cette connaissance de l'emplacement des exploitations échantillonnées autorise l'utilisation des méthodes de lissage spatiale décrites plus loin dans le présent article. Traditionnellement, les estimations du AAGIS ont été présentées uniquement sous forme de tableaux de moyennes

Les cartes constituent une méthode pratique de présentation des données pour un certain nombre de raisons. Premièrement, les données présentées sous cette forme s'interprètent rapidement; confronté à un trop grand nombre de tableaux, le client risque de négliger certaines variations locales ou de se sentir «submergé» par les chiffres. Deuxièmement, grâce aux cartes, le client peut plus facilement établir un lien entre

La cartographie des variations régionales du rendement économique des exploitations étudiées constitue une forme très efficace de présentation spatiale des informations. Nous utilisons une méthode de régression non paramétrique pour réaliser le lissage spatial des données d'enquête provenant de chacune des exploitations, pour les présenter ensuite sous forme de cartes. Les améliorations récentes apportées à la puissance d'analyse et la disponibilité de systèmes SIG abordables et de haute qualité ont fait de cette forme de présentation une solution de rechange pratique à la méthode traditionnelle de présentation des résultats d'enquête sous forme de tableaux.

calculées pour toute l'Australie, pour chaque Etat et pour chaque industrie à l'intérieur des Etats. Toutefois, les pré-occupations suscitées au sein de l'industrie rurale et des gouvernements par l'incidence combinée des sécheresses qui ont sévi dans certaines régions de l'Australie et du déclin des prix de certains produits ont mis en lumière l'importance de recueillir des informations plus pertinentes et plus détaillées sur les tendances régionales affichées par le rendement des exploitations. En particulier, on a relevé la nécessité d'obtenir des informations sur la distribution spatiale du rendement des exploitations qui reflèteraient la variabilité réelle du climat et de la production d'un bout à l'autre du territoire australien.

<sup>1</sup> Ann Cowling, CSIRO Division of Fisheries, GPO Box 1338, Hobart TAS 7001, Australie et Australian Bureau of Agricultural and Resource Economics; Ray Chambers, Department of Social Statistics, University of Southampton, Highfield, Southampton SO17 1BJ, Royaume-Uni; Ray Lindsay et Bhamathy Parameswaran, Australian Bureau of Agricultural and Resource Economics, GPO Box 1563, Canberra ACT 2601, Australie.





## ANNEXE II

## Preuve du Théorème 4

- Pour le plan de Chao, il suffit d'utiliser (6), (9) et (15).
- Pour le plan systématique randomisé, il suffit d'utiliser l'approximation des  $\pi^{(N;t,j)}$  donnée par Deville (p. 21)

$$\pi^{(N;t,j)} \approx \pi^{(N;t)} \pi^{(N;j)} \frac{n - 1}{n - \pi^{(N;t)} - \pi^{(N;j)}}. \quad (22)$$

Cette expression est obtenue à partir de l'hypothèse

$$\text{Max}_{1 \leq i \leq N} \left\{ \frac{\pi^{(N;i)}}{n} \right\} \rightarrow 0.$$

Cette dernière hypothèse est vérifiée car  $n \rightarrow \infty$ .

- Pour le plan réjectif, par le résultat d'Hájek (1964, p. 1508), nous avons

$$-\Delta_{(N;t,j)}^{(N;t,j)} \approx \frac{[1 - \pi^{(N;t)}][1 - \pi^{(N;j)}]}{d - [1 - \pi^{(N;t)}][1 - \pi^{(N;j)}]} \quad (23)$$

pour  $d \rightarrow \infty$ . Notons que (23) reste valable pour le plan de Rao-Sampford (cf. Hájek 1981, Théorème 8.2, p. 82). En utilisant l'approximation (Hájek 1964, p. 1521),

$$\{d - [1 - \pi^{(N;t)}][1 - \pi^{(N;j)}]\}^{-1} \approx \frac{d(n-1)}{n},$$

nous obtenons le résultat du théorème.

Cqfd.

## REMERCIEMENTS

L'auteur tient à remercier les arbitres qui ont émis de nombreux commentaires constructifs, permettant d'améliorer considérablement cet article.

## BIBLIOGRAPHIE

- BETHLEHEM, J.G., et SCHUEERHOFF, H. (1984). Second-order inclusion probabilities in sequential sampling without replacement with unequal probabilities. *Biometrika*, 71, 642-644.
- CHAO, M.T. (1982). A general purpose unequal probability sampling plan. *Biometrika*, 69, 653-656.
- DEVILLE, J.-C. (Sans date). Cours de sondage, Chapitre III: les outils de bases. Polycopié de l'ENSAE, Paris.
- HÁJEK, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*, 35, 1491-1523.
- HÁJEK, J. (1981). *Sampling from a Finite Population*. New York et Bassel: Marcel Dekker, Inc.
- HARTLEY, H.O., et RAO, J.N.K. (1962). Sampling with unequal probabilities without replacement. *Annals of Mathematical Statistics*, 33, 350-374.
- HORVITZ, D.G., et THOMPSON, D.J. (1951). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.
- MCLEOD, A.I., et BELLHOUSE, D.R. (1983). A convenient algorithm for drawing a simple random sample. *Applied Statistics*, 32, 2.
- RAO, J.N.K. (1965). On two simple schemes of unequal probability sampling without replacement. *Journal of the Indian Statistical Association*, 3, 173-180.
- SAMPFORD, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54, 494-513.
- YATES, F., et GRUNDY, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society, série B*, 1, 253-261.

$$q_{(j)}^* = \begin{cases} \pi_{(N-j-1;j)}^* \pi_{(j;j)} \left(1 - \frac{n}{1}\right) & \text{si } j > n+1; \\ \pi_{(n+1;j)}^* + \pi_{(n+1;j)} - 1 & \text{si } j \leq n+1; \end{cases}$$

et  $a_i^*$  est défini par (17).

Maintenant, avec ces deux lemmes, nous pouvons démontrer le théorème 3.

### Preuve du théorème 3

**Cas 1:** Si  $j > n+1$ , par le lemme 2, nous avons

$$\pi_{(N;j)}^* = \pi_{(j-1;j)}^* \pi_{(j;j)} \left(1 - \frac{n}{1}\right) \prod_{\ell=j+1}^N \left[1 - \pi_{(\ell;j)} \frac{n}{2}\right].$$

Par le lemme 1, cette dernière expression devient

$$\pi_{(N;j)}^* = p_{(j)}^* \pi_{(j;j)} \left(1 - \frac{n}{1}\right) \prod_{\ell=1}^{j-1} \left(\frac{n}{1}\right)$$

$$\left[1 - \pi_{(\ell;j)} \frac{n}{1}\right] \prod_{\ell=j+1}^{b=j+1} \left[1 - \pi_{(\ell;j)} \frac{n}{2}\right].$$

En multipliant cette dernière expression par

$$\left[ \frac{1 - \pi_{(j;j)} \frac{n}{1}}{1 - \pi_{(j;j)} \frac{n}{2}} \prod_{\ell=j+1}^N \frac{1 - \pi_{(\ell;j)} \frac{n}{1}}{1 - \pi_{(\ell;j)} \frac{n}{2}} \right] \left[ \frac{1 - \pi_{(\ell;j)} \frac{n}{1}}{1 - \pi_{(\ell;j)} \frac{n}{2}} \right] = 1,$$

et en regroupant certains termes, nous obtenons

$$\pi_{(N;j)}^* = \pi_{(j;j)}^* p_{(j)}^* \left[ \frac{n-1}{n} \right] \prod_{\ell=a_i^*}^j \left[ \frac{n-1}{n} \right]$$

$$\left[ 1 - \pi_{(\ell;j)} \frac{n}{1} \right] \prod_{\ell=j+1}^N \left[ \frac{1 - \pi_{(\ell;j)} \frac{n}{2}}{1 - \pi_{(\ell;j)} \frac{n}{1}} \right] \left[ \frac{1 - \pi_{(b;j)} \frac{n}{2}}{1 - \pi_{(b;j)} \frac{n}{1}} \right].$$

Par le lemme 1, cette dernière expression devient

$$\pi_{(N;j)}^* = \left[ \frac{n-1}{n} \right] \pi_{(j;j)}^* \pi_{(j;j)} \prod_{\ell=j+1}^N \pi_{(\ell;j)} \left[ \frac{1 - \pi_{(\ell;j)} \frac{n}{2}}{1 - \pi_{(\ell;j)} \frac{n}{1}} \right]. \quad (18)$$

Par le lemme 1, nous obtenons finalement

$$\pi_{(N;j)}^* \approx \pi_{(N;j)} \pi_{(N;j)} \frac{\pi_{(n+1;j)}}{\pi_{(n+1;j)} + \pi_{(n+1;j)}}. \quad (21)$$

Cqfd.

Si  $n$  est suffisamment grand

$$\frac{1 - \pi_{(\ell;j)} \frac{n}{2}}{1 - \pi_{(\ell;j)} \frac{n}{1}} \approx \left[ 1 - \pi_{(\ell;j)} \frac{n}{2} \right] \left[ 1 + \pi_{(\ell;j)} \frac{n}{1} \right];$$

$$\approx 1 + \frac{n}{2\pi_{(\ell;j)}} - \frac{n}{2\pi_{(\ell;j)}^2};$$

(19)

Dès lors (18) devient,

$$\pi_{(N;j)}^* \approx \left[ \frac{n-1}{n} \right] \pi_{(j;j)}^* \pi_{(j;j)} \prod_{\ell=j+1}^N \left[ 1 - \pi_{(\ell;j)} \frac{n}{1} \right]. \quad (20)$$

Finalement, par le lemme 1, cette dernière expression s'écrit:

$$\pi_{(N;j)}^* \approx \pi_{(N;j)} \frac{n-1}{n} \pi_{(j;j)}.$$

**Cas 2:** Si  $j \leq n+1$ , le lemme 2 donne

$$\pi_{(N;j)}^* = [\pi_{(n+1;j)} + \pi_{(n+1;j)} - 1] \prod_{\ell=n+2}^N \left[ 1 - \pi_{(\ell;j)} \frac{n}{2} \right];$$

c'est-à-dire

$$\pi_{(N;j)}^* = \prod_{\ell=n+2}^N \left[ 1 - \pi_{(\ell;j)} \frac{n}{1} \right] \prod_{\ell=b=n+2}^N \frac{n}{1}$$

$$\left[ \frac{1 - \pi_{(b;j)} \frac{n}{2}}{1 - \pi_{(b;j)} \frac{n}{1}} \right] \left[ \pi_{(n+1;j)} + \pi_{(n+1;j)} - 1 \right].$$

En utilisant l'approximation (19), nous obtenons

$$\pi_{(N;j)}^* \approx \left\{ \prod_{\ell=n+2}^N \left[ 1 - \pi_{(\ell;j)} \frac{n}{1} \right] \right\}_2$$

$$\pi_{(n+1;j)} \frac{\pi_{(n+1;j)}}{\pi_{(n+1;j)} + \pi_{(n+1;j)}}.$$



Nous avons une erreur moyenne de -0.006999 avec un écart type de 0.006438. Le centre de gravité du nuage de points se trouve en (0.02957; 0.036606). Nous aboutissons à la même conclusion que celle de l'exemple 1. Le second exemple conduit à de plus mauvaises approximations. Cela est simplement dû aux grandes probabilités d'inclusion d'ordre un.

7. CONCLUSION

Le plan de Chao possède plusieurs avantages: (i) il est séquentiel, (ii) les probabilités d'inclusion d'ordre deux sont positives et (iii) la variance de Yates-Grundy est toujours positive. Par contre, les probabilités d'inclusion d'ordre deux sont difficiles à calculer. C'est pourquoi, nous proposons de les approximer. Nous avons constaté que cette approximation est d'autant meilleure que le début de la population est constitué d'unités ayant de petites  $\pi^{(N;i)}$  et que la fin de la population est constituée d'unités ayant de grandes  $\pi^{(N;i)}$ . Nous avons comparé notre approximation avec d'autres approximations données pour le plan systématique randomisé, le plan réjectif et le plan de Rao-Sampford. Nous avons conclu que ces approximations sont équivalentes si les probabilités d'inclusion d'ordre un sont petites et si la taille de l'échantillon est grande. Les deux exemples numériques qui terminent cet article, confirment les bons résultats de notre approximation.

ANNEXE I

Preuve du théorème 3

Avant d'entamer la preuve de ce théorème, on peut démontrer les deux lemmes suivants.

Lemme 1

$$\pi^{(k;i)} = p_*^{(i)} \prod_{j=1}^k \left[ 1 - \pi^{(i;j)} \right] \frac{1}{n};$$

où

$$p_*^{(i)} = \begin{cases} \pi^{(i;i)} & \text{si } i > n + 1; \\ \pi^{(n+1;i)} & \text{si } i \leq n + 1; \end{cases}$$

$$a_*^i = \begin{cases} i + 1 & \text{si } i > n + 1; \\ n + 2 & \text{si } i \leq n + 1. \end{cases}$$

Lemme 2

$$\pi^{(k;i)} = q_*^{(f)} \prod_{j=1}^k \left[ 1 - \pi^{(i;j)} \right] \frac{1}{2};$$

où  $i < j_i$

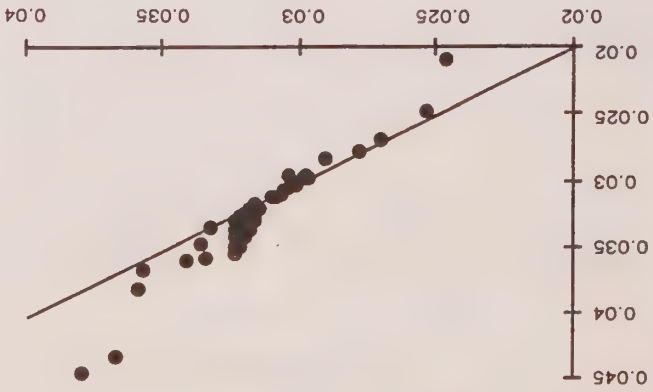


Figure 2. Approximations et vraies valeurs de  $-\Delta^{(N;f)}/\pi^{(N;f)}$ , dans le cas de l'exemple 1

ces couples ont une grande probabilité de se trouver dans l'échantillon étant donné que  $\pi^{(N;j)}$  est grand. Donc notre variance approchée (10) est tout à fait acceptable.

Exemple 2

Les probabilités d'inclusion d'ordre un sont données dans la figure 3. Ici nous constatons que ces probabilités sont plus dispersées que dans l'exemple 1. La figure 4 donne les vraies valeurs ainsi que les approximations de  $-\Delta^{(N;f)}/\pi^{(N;f)}$ .

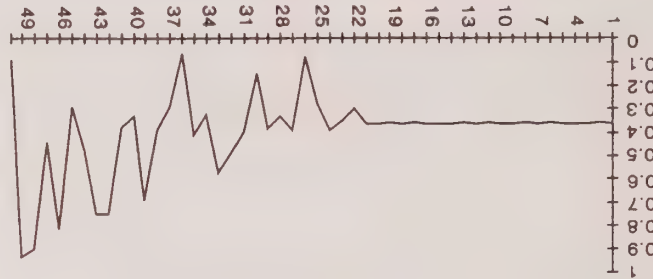


Figure 3. Probabilités d'inclusion d'ordre un, dans le cas de l'exemple 2

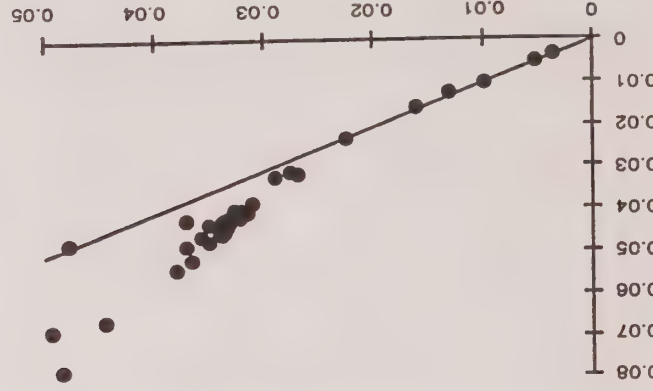


Figure 4. Approximations et vraies valeurs de  $-\Delta^{(N;f)}/\pi^{(N;f)}$ , dans le cas de l'exemple 2

## 5. COMPARAISON AVEC D'AUTRES PLANS

Au lieu de comparer les probabilités d'inclusion d'ordre deux, nous allons comparer les quantités  $-\Delta_{(N;i,j)}^{(N;i,j)}/\pi_{(N;i,j)}^{(N;i,j)}$  qui sont utiles pour calculer la variance de Yates-Grundy. Nous allons regarder ce que donnent ces quantités pour le plan de Chao, le plan systématique randomisé (Hartley et Rao 1962), le plan réjectif (Hájek 1964) et le plan de Rao-Sampford (Rao 1965 et Sampford 1967).

### Théorème 4

$$\left\{ \begin{array}{l} \frac{-\Delta_{(N;i,j)}^{(N;i,j)}}{\pi_{(N;i,j)}^{(N;i,j)}} \approx \left\{ \begin{array}{ll} \frac{1 - \pi_{(N;j)}^{(N;j)}}{n - 1}, & \text{pour le plan de Chao;} \\ \frac{1 - \pi_{(N;i)}^{(N;i)} - \pi_{(N;j)}^{(N;j)}}{n - 1}, & \text{pour le plan systématique randomisé;} \\ \frac{n [1 - \pi_{(N;i)}^{(N;i)}] [1 - \pi_{(N;j)}^{(N;j)}]}{d(n - 1)}, & \text{pour le plan réjectif et pour le plan de Rao-Sampford.} \end{array} \right. \end{array} \right.$$

La preuve de ce théorème se trouve en appendice II.

Il est important de noter que l'approximation proposée pour le plan systématique randomisé provient de l'approximation de Deville (p. 21) et non de la célèbre approximation de Hartley-Rao (1962). Nous n'avons pas pu utiliser la formule de Hartley-Rao étant donné que celle-ci repose sur l'hypothèse asymptotique,  $n$  fixé et  $N \rightarrow \infty$ , différente de celle adoptée dans cet article.

Nous constatons que si les  $\pi_{(N;i,j)}^{(N;i,j)}$  sont petites,  $-\Delta_{(N;i,j)}^{(N;i,j)}/\pi_{(N;i,j)}^{(N;i,j)}$  est équivalent pour le plan de Chao et pour le plan systématique. Mais, nous constatons que  $-\Delta_{(N;i,j)}^{(N;i,j)}/\pi_{(N;i,j)}^{(N;i,j)}$  est toujours plus petit dans le cas systématique que dans le cas de Chao. Cela est certainement dû au fait que l'approximation pour le plan systématique sous-estime  $-\Delta_{(N;i,j)}^{(N;i,j)}/\pi_{(N;i,j)}^{(N;i,j)}$ . On peut s'en rendre compte en remplaçant  $\pi_{(N;i)}^{(N;i)}$  et  $\pi_{(N;j)}^{(N;j)}$  par  $n/N$ . On obtient alors

$$-\frac{\Delta_{(N;i,j)}^{(N;i,j)}}{\pi_{(N;i,j)}^{(N;i,j)}} \approx \frac{N(n - 1)}{N - 2n}$$

pour le plan systématique randomisé. Or, un tel plan est équivalent au plan simple. Il faudrait dès lors que,

$$-\frac{\Delta_{(N;i,j)}^{(N;i,j)}}{\pi_{(N;i,j)}^{(N;i,j)}} = \frac{N(n - 1)}{N - n}.$$

Nous proposons de redresser l'approximation de  $-\Delta_{(N;i,j)}^{(N;i,j)}/\pi_{(N;j)}^{(N;j)}$  pour le plan systématique en la multipliant par

$$\frac{N - n}{N - 2n} = \frac{1}{1 - 2f},$$

où  $f = n/N$  est le taux de sondage.

L'approximation de  $-\Delta_{(N;i,j)}^{(N;i,j)}/\pi_{(N;i,j)}^{(N;i,j)}$  pour le plan de Chao est également du même ordre que celle du plan réjectif. En effet, si les  $\pi_{(N;i)}^{(N;i)}$  sont petites, nous avons l'approximation

## 6. EXEMPLES NUMÉRIQUES

Les deux exemples suivants correspondent à deux cas extrêmes. Dans le premier exemple, les  $\pi_{(N;i)}^{(N;i)}$  sont peu dispersées; dans le second, elles le sont beaucoup plus. Prenons un petit échantillon de taille 20. La population aura une taille de 50 pour que les  $\pi_{(N;i)}^{(N;i)}$  ne soient pas trop petites. Nous nous sommes volontairement mis dans de mauvaises conditions, pour montrer que même avec un échantillon de taille 20 et une petite population, les résultats asymptotiques constituent déjà une bonne approximation.

### Exemple 1

Prenons les probabilités d'inclusion d'ordre un représentées par la figure 1.

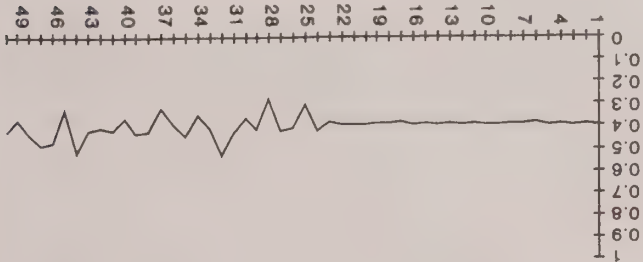


Figure 1. Probabilités d'inclusion d'ordre un, dans le cas de l'exemple 1

La figure 2 représente en ordonnée les vraies valeurs de  $-\Delta_{(N;i,j)}^{(N;i,j)}/\pi_{(N;i,j)}^{(N;i,j)}$  pour le plan de Chao et en abscisse les approximations. Nous avons également représenté la droite où les approximations sont égales aux vraies valeurs. Les approximations seront d'autant meilleures que les points sont proches de la droite. Nous avons une erreur moyenne de  $-0.000569$  avec un écart type de 0.0015996. Ceci est très faible par rapport à l'ordre de grandeur des approximations. Le centre de gravité du nuage de point se trouve en  $(0.0313; 0.0318)$ . Il peut paraître surprenant qu'il y ait moins de points à gauche du centre de gravité qu'à droite. Cela est simplement dû au fait que la plupart des points à gauche du centre de gravité sont superposés.

Nous constatons que les couples  $(i,j)$  avec  $i < j$  tel que  $\pi_{(N;j)}^{(N;j)}$  est grand correspondent à des points se trouvant à gauche. Ce sont les couples ayant la meilleure approximation. De plus,



## 4. ESTIMATEUR DE VARIANCE

La relation (7) conduit à l'approximation suivante des  $\Delta_{(N;i,j)}$ :

$$\Delta_{(N;i,j)} = \pi_{(N;i)} \pi_{(N;j)} \frac{u}{d_{(j)} - 1}, \quad \text{si } i < j. \quad (9)$$

(2), (7) et (9) permettent d'établir une expression asymptotique pour l'estimateur de Yates-Grundy.

$$V^c = \frac{1}{n-1} \sum_{j \in S_N} [1 - d_{(j)}] \sum_{i \in S_N; i < j} \left[ \frac{\pi_{(N;i)}}{Y_i} - \frac{\pi_{(N;j)}}{Y_j} \right]^2. \quad (10)$$

Mais cette expression a tendance à sous-estimer la variance. En effet, pour établir la relation (6), on utilise l'approximation (19) de l'appendice I. Cette approximation implique toujours que:

$$\pi_{(N;i,j)} > \pi_{(N;i)} \pi_{(N;j)} \frac{u}{n-1} d_{(j)}. \quad (11)$$

On peut s'en rendre compte facilement en constatant que (20) est obtenu à partir de (18) en utilisant l'approximation (19). L'inégalité (11) est donc vraie pour  $j > n+1$ . Pour  $j \leq n+1$ , il suffit de constater que (21) est également obtenu à partir de (19). L'inégalité (11) implique que:

$$-\Delta_{(N;i,j)} \frac{\pi_{(N;i,j)}}{1 - d_{(j)}} > \frac{\pi_{(N;i,j)}}{1 - d_{(j)}}, \quad (12)$$

étant donnée que  $\Delta_{(N;i,j)} > 0$ . Par (2), (10) et (12), nous avons effectivement

$$V^c > V.$$

Pour remédier à ce problème de sous-estimation de la variance, nous proposons de faire un redressement sur (9). Il est bien connu que:

$$\sum_N \pi_{(N;i,j)} = (n-1) \pi_{(N;j)}. \quad (13)$$

L'approximation (7) ne respecte pas la contrainte (13). Le redressement consiste à supposer les  $d_{(j)}$  inconnus et à les choisir de manière à ce que (13) soit satisfaite pour l'approximation des probabilité d'ordre deux. C'est-à-dire:

$$\sum_{j=1}^i \pi_{(N;i,j)} \pi_{(N;j)} \frac{u}{n-1} d_{(j)} + \sum_{j=i+1}^N \pi_{(N;i,j)} \pi_{(N;j)} \frac{u}{n-1} d_{(i)} = (n-1) \pi_{(N;j)}.$$

Cette contrainte peut s'écrire:

$$\sum_{j=1}^i \pi_{(N;i,j)} + \sum_N^{j=i+1} \pi_{(N;i,j)} \frac{u}{n-1} d_{(i)} = n - d_{(i)}. \quad (14)$$

Étant donné  $\sum_N^{j=1} \pi_{(N;j)} = n$ , la contrainte (14) est pratiquement vérifiée si

$$(15)$$

$$\sum_N \pi_{(N;i)} \frac{u}{n - \pi_{(N;j)}} \approx \sum_N^{j=i+1} \pi_{(N;j)}. \quad (16)$$

La relation (16) est plausible étant donné que la différence entre la partie gauche et la partie droite de (16) a comme

borne inférieure

$$\frac{1}{N} \sum_N^{j=i+1} \pi_{(N;i)} [\pi_{(N;i)} - \pi_{(N;j)}],$$

et comme borne supérieure

$$\frac{1}{N} \sum_N^{j=i+1} \pi_{(N;i)} [\pi_{(N;i)} - \pi_{(N;j)}].$$

Ces deux bornes sont proches de zéro lorsque les  $\pi_{(N;j)}$  sont peu dispersées. Ceci signifie que la solution (15) est approchée lorsque les  $\pi_{(N;j)}$  sont petits. Ces deux bornes sont également d'autant plus proches de zéro que  $j$  est grand. Donc la solution (15) vérifie d'autant plus (13) que  $j$  est grand. Ceci implique que notre approximation (9) est très bonne pour les couples d'unités  $(i, j)$  ( $i < j$ ) tels que l'unité  $j$  est située à la fin de la population. En fait, nous voulons que l'approximation (9) soit la meilleure pour les couples d'unités  $(i, j)$  ayant une grande probabilité de se trouver dans l'échantillon (c'est-à-dire, pour les couples  $(i, j)$  ( $i < j$ ) dont  $\pi_{(N;j)}$  est la plus grande). Il est donc préférable de placer les unités possédant des grandes probabilités d'inclusion d'ordre un à la fin de la population.

Si on choisit de prendre  $d_{(i)} = \pi_{(N;i)}$ , nous avons des  $d_{(i)}$  plus petit que (8). Ceci conduit à une approximation de la variance plus grande. Cette solution est d'autant plus acceptable qu'elle correspond au résultat du plan simple sans remise. En effet, si on remplace dans (7)  $\pi_{(N;i,j)}$  et  $d_{(j)}$  par  $n/N$ , on obtient

$$\pi_{(N;i,j)} \approx \frac{n(N-n-1)}{n(n-1)}, \quad \text{si } i > n+1.$$

Cette expression correspond, bien évidemment, au résultat du plan simple sans remise.

En conclusion, nous approximations  $\Delta_{(N;i,j)}$  par (9) avec  $d_{(i)} = \pi_{(N;i)}$ . Nous supposons que la population est tirée de manière à ce que les unités ayant des petites  $\pi_{(N;j)}$  se situent au début de la population et que les unités ayant des grandes  $\pi_{(N;j)}$  se situent à la fin de la population. Nous supposons également que les  $\pi_{(N;j)}$  ne sont pas trop dispersées pour les  $n+1$  premières unités de la population.

Comme  $\pi^{(i,j)} < 1$  pour tous  $i$  et  $j$  tels que  $i \leq \ell \leq k$ , cette condition est toujours satisfaite. Donc, dans le cadre de cet article, nous n'aurons jamais de probabilités d'inclusion d'ordre deux nulles.

De plus, la quantité  $\Delta^{(N,i,j)}$  est toujours négative si on utilise le plan de Chao (Chao 1982, p. 656). Dès lors, la variance de Yates-Grundy a l'avantage d'être toujours positive.

### 3. APPROXIMATION DES PROBABILITÉS D'INCLUSION D'ORDRE 2

Le théorème suivant nous donne une expression asymptotique pour les probabilités d'inclusion d'ordre deux, pour le plan de Chao.

#### Théorème 3

$$\pi^{(N,i,j)} \approx \begin{cases} \frac{n-1}{n} d^{(j)} \pi^{(N,j)} & , \text{ si } j > n+1; \\ \frac{\pi^{(N,i)} \pi^{(N,j)} + \pi^{(n+1,j)} \pi^{(n+1,i)}}{\pi^{(N,i)} \pi^{(N,j)} + \pi^{(n+1,j)} \pi^{(n+1,i)}} & , \text{ si } j \leq n+1; \end{cases} \quad (6)$$

où  $p^{(j)} = \pi^{(j,j)}$  et  $i < j$ .

La preuve de ce théorème se trouve en appendice I.

Nous constatons que cette approximation a une structure différente suivant que  $j > n+1$  ou que  $j \leq n+1$ . Pour éviter la variable auxiliaire, de manière à ce qu'il y ait équivalence entre ces deux structures. Considérons l'hypothèse énoncée dans l'introduction, selon laquelle les valeurs de la variable auxiliaire sont peu dispersées pour les  $n+1$  premières unités de la population. Plus précisément, nous supposons que la variable auxiliaire est constante pour les  $n+1$  premières unités, c'est-à-dire:

$$\pi^{(n+1,i)} = \frac{n+1}{n} \quad \text{pour } i \leq n+1.$$

Dans ce cas,

$$\frac{\pi^{(n+1,i)} + \pi^{(n+1,j)}}{n-1} = \frac{\pi^{(n+1,i)} \pi^{(n+1,j)}}{n-1}.$$

En utilisant (6), nous avons l'approximation suivante pour les probabilités d'inclusion d'ordre deux

$$\pi^{(N,i,j)} \approx \pi^{(N,i)} \pi^{(N,j)} \frac{n-1}{n} d^{(j)} \quad \text{si } i < j; \quad (7)$$

ou

$$p^{(j)} = \begin{cases} \pi^{(j,j)} & , \text{ si } j > n+1, \\ \pi^{(n+1,j)} & , \text{ si } j \leq n+1. \end{cases} \quad (8)$$

hasard dans  $S_k$ . La procédure démarre à partir d'un échantillon initial,  $S_n = U_n$ , constitué des  $n$  premières unités de la population.

Le plan de Chao a l'avantage d'être séquentiel. En effet, il permet de sélectionner un échantillon par un simple parcours séquentiel de la population. Le plan systématique est un autre plan séquentiel souvent utilisé. Mais ce dernier a le désavantage d'induire des probabilités d'inclusion d'ordre deux nulles. On peut éviter ce problème en randomisant le plan systématique. Dans ce cas, la population est triée de manière aléatoire avant que l'échantillon ne soit sélectionné. Cette opération élimine partiellement le problème des probabilités d'inclusion d'ordre deux nulles. Par contre, comme nous le verrons à la fin de cette section, le plan de Chao a l'avantage de ne pas avoir de probabilités d'inclusion d'ordre deux nulles.

Une randomisation n'est donc pas nécessaire pour ce dernier. Le plan réjéctif et le plan de Rao-Sampford ont le désavantage de ne pas être séquentiels. En effet, les unités sont sélectionnées au hasard avec remise dans la population. Si une unité est sélectionnée deux fois, on est obligé de sélectionner un nouvel échantillon. Ces deux plans, bien que plus simple à comprendre, sont plus difficiles à mettre en oeuvre que le plan de Chao.

Le théorème suivant, qui est une application directe du théorème donné par Chao (1982), donne une relation entre la probabilité d'inclusion d'ordre un,  $\pi^{(k,i)}$ , de la  $i$ -ième unité de  $U_k$  et la probabilité d'inclusion d'ordre un,  $\pi^{(k+1,i)}$ , de la  $i$ -ième

#### Théorème 1

$$\pi^{(k+1,i)} = \begin{cases} [1 - \pi^{(k+1,k+1)}] R^{(k,i)} \pi^{(k,i)} & , \text{ pour } i < k+1; \\ \pi^{(k+1,k+1)} & , \text{ pour } i = k+1; \end{cases} \quad (4)$$

où

$$R^{(k,i)} = \begin{cases} \frac{1 - \pi^{(n+1,i)}}{\pi^{(n+1,n+1)}} & , \text{ pour } k = n, \\ \frac{1}{n} & , \text{ pour } k \geq n+1. \end{cases} \quad (5)$$

Les probabilités d'inclusion d'ordre deux peuvent être calculées de manière itérative en utilisant le théorème suivant:

#### Théorème 2 (Chao, 1982)

$$\pi^{(k,i,j)} =$$

$$\begin{cases} \{1 - \pi^{(k,k)} [R^{(k-1,i)} + R^{(k-1,j)}] \pi^{(k-1,i,j)}\} \pi^{(k-1,i,j)} & , \text{ pour } i < j < k; \\ \pi^{(k,k)} [1 - R^{(k-1,i)}] \pi^{(k-1,i)} & , \text{ pour } i < j = k. \end{cases}$$

Bethlehem et Schuerhoff (1984) donnent une condition nécessaire et suffisante pour que les probabilités d'inclusion d'ordre deux soient strictement positives, pour une population  $U_k$ :

$$\# \{i : i \leq \ell \text{ et } \pi^{(i,j)} = 1\} \neq n-1, \text{ pour tout } \ell \text{ tel que } n < \ell \leq k.$$



# Variance asymptotique pour un plan séquentiel sans remise à probabilités inégales

YVES G. BERGER<sup>1</sup>

## RÉSUMÉ

Nous proposons une approximation des probabilités d'inclusion d'ordre deux pour le plan de Chao (1982) en vue d'obtenir un estimateur de variance approché pour l'estimateur de Horvitz et Thompson. Ensuite, nous comparons cette variance à d'autres approximations données pour le plan systématique randomisé (Hartley et Rao 1962), le plan réjectif (Hájek 1964) et le plan de Rao-Sampford (Rao 1965 et Sampford 1967). Nous concluons que ces approximations sont équivalentes si les probabilités d'inclusion d'ordre un sont petites et si la taille de l'échantillon est grande.

**MOTS CLÉS :** Sondage avec remise; plan systématique randomisé; plan réjectif; plan de Rao-Sampford; probabilités d'inclusion; Horvitz-Thompson; Yates-Grundy.

## 1. INTRODUCTION

Soit  $U_N$  une population finie contenant  $N$  unités. Considérons  $U_k$  un sous-ensemble de  $U_N$  constitué des  $k$  premières unités de  $U_N$ . Nous notons  $\pi_{(k;t)}$  les probabilités d'inclusion d'ordre un pour une population  $U_k$ . Nous supposons qu'elles sont proportionnelles à une variable auxiliaire. Ces probabilités ont deux arguments; la taille  $k$  de la population et le numéro d'ordre  $i$  de l'unité dans la population. Nous supposons que  $\pi_{(k;t)} < 1$  pour tous les  $i$  et tous les  $k > n$ . Cette hypothèse a plus de chance d'être violée dans le cas où  $k$  est petit, c'est-à-dire proche de  $n$ . Nous pouvons remédier à ce problème en supposant que les valeurs de la variable auxiliaire sont peu dispersées pour les unités se trouvant au début de la population.

Nous notons  $\pi_{(k;t)}$  la probabilité d'inclusion d'ordre deux des unités  $i$  et  $j$  pour une population  $U_k$ . Ces probabilités dépendent du plan de sondage utilisé.

Nous utiliserons l'estimateur de Horvitz-Thompson (1951) pour estimer le total  $\sum_{i=1}^N Y_i$  d'une variable  $Y$ . Cet estimateur est donné par

$$t_{HT} = \sum_{i \in S_N} \frac{\pi_{(N;i)}}{Y_i}; \quad (1)$$

où  $S_N$  est un échantillon de  $U_N$ . Nous supposons que la taille de  $S_N$  est constante et égale à  $n$ . Étant donné que la taille de l'échantillon est fixée, un estimateur de la variance de (1) est donné par l'estimateur de Yates-Grundy (1953),

$$v = \sum_{j \in S_N} \sum_{i \in S_N; i < j} \left[ \frac{\pi_{(N;i,j)}}{\pi_{(N;i)}} \frac{\pi_{(N;j)}}{Y_i} - \frac{\pi_{(N;j)}}{Y_j} \right]^2, \quad (2)$$

où

$$\Delta_{(N;i,j)} = \pi_{(N;i,j)} - \pi_{(N;i)}\pi_{(N;j)}. \quad (3)$$

Considérons la séquence de taille d'échantillon  $\{n_1, n_2, \dots, n_v, \dots\}$  et la séquence de taille de population  $\{N_1, N_2, \dots, N_v, \dots\}$ , où  $n_v$  et  $N_v$  augmentent lorsque  $v \rightarrow \infty$ . Pour simplifier le problème nous supprimons l'indice  $v$ . L'approche asymptotique adoptée ici est celle d'Hájek (1964):

$$d = \sum_N \pi_{(N;j)} [1 - \pi_{(N;j)}] \rightarrow \infty, \quad (4)$$

ce qui signifie que  $n \rightarrow \infty$  et  $(N - n) \rightarrow \infty$ , étant donné que  $d \leq \sum_{j=1}^N [1 - \pi_{(N;j)}] = N - n$  et  $d \leq \sum_{j=1}^N \pi_{(N;j)} = n$ .

Dans la section 2, nous présentons le plan de sondage de Chao (1982) ainsi que trois résultats se rapportant aux probabilités d'inclusion d'ordre un et deux. Dans la section 3, nous donnons une approximation des  $\pi_{(N;i,j)}$ . Dans la section 4, nous proposons une approximation de la variance de Yates-Grundy. La section 5 est consacrée à la comparaison de cette approximation de variance avec d'autres approximations proposées pour le plan systématique randomisé, le plan réjectif et le plan de Rao-Sampford. Deux exemples numériques sont présentés dans la section 6.

## 2. PLAN DE SONDAGE DE CHAO

Il s'agit d'un plan de sondage sans remise à probabilités inégales, à taille fixée. Cette méthode est la généralisation de la méthode de McLeod et Bellhouse (1983) pour un plan simple.

Soit  $S_k$  un échantillon de taille  $n$  de  $U_k$  avec un ensemble  $\{\pi_{(k;t)}; i \in U_k\}$  de probabilités d'inclusion d'ordre un. Le plan de Chao permet d'avoir un échantillon  $S_{k+1}$  de taille  $n$  de  $U_{k+1}$  avec un ensemble  $\{\pi_{(k+1;t)}; i \in U_{k+1}\}$  de probabilités d'inclusion d'ordre un. La méthode consiste à sélectionner la  $(k+1)$ -ième unité avec une probabilité  $\pi_{(k+1;k+1)}$ . Si cette unité  $n$  est pas sélectionnée, alors nous prenons  $S_{k+1} = S_k$ ; sinon nous prenons  $S_{k+1} = S_k \cup \{k+1\}$ , où  $j$  est une unité sélectionnée au

- NATIONAL CENTER FOR HEALTH STATISTICS (1996). National Health and Nutrition Examination Survey III Report (à paraître). National Center for Health Statistics, Hyattsville, MD.
- RUST, K., et KALTON, G. (1987). Strategies for collapsing strata for variance estimation. *Journal of Official Statistics*, 3, 69-81.
- SATTEIRTHWAITE, F.E. (1941). Synthesis of variance. *Psychometrika*, 6, 309-316.
- SATTEIRTHWAITE, F.E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2, 110-114.
- WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.



**Tableau 5**  
Estimation de  $V_B$  et  $R_{wv}$  pour les douze variables de la NHANES III avec erreur-type et apport relatif de la variance intra-UPF

Variable	$V_B$	$ei(V_B)$	$V(V_B)^{-1}V(V_w)$
HAE2	0.0000126	0.0000188	0.020
HAET	0.0000532	0.0000445	0.030
HAD1	-0.00000208	0.00000246	0.186
HAR3	0.0000825	0.0000703	0.047
BMPTH	0.0193	0.0114	0.027
BMPWT	0.0174	0.0400	0.096
HDRESULT	0.0887	0.0744	0.010
TCRESULT	0.270	0.253	0.031
LEAD	0.00269	0.00188	0.168
log(LEAD)	0.000468	0.000205	0.012
BPIK1	1.823	0.997	0.081
BPIK5	-0.0351	0.0793	0.367

$R_{wv}$	$ei(R_{wv})$	$V(R_{wv})^{-1}V_w R_{wv}^2 V(V_w)$
HAE2	1.327	0.491
HAET	1.648	0.556
HAD1	0.783	0.247
HAR3	1.676	0.600
BMPTH	1.864	0.530
BMPWT	1.168	0.391
HDRESULT	2.193	1.020
TCRESULT	1.458	0.436
LEAD	1.694	0.555
log(LEAD)	3.221	1.025
BPIK1	2.699	1.142
BPIK5	0.861	0.300

de  $V(V_w)$ . Relativement parlant,  $V(V_w)$  n'est pas aussi élevée pour BPIK5, mais la proportion  $V(R_{wv})^{-1}V_w R_{wv}^2 V(V_w)$  demeure appréciable, car  $V_w$  n'est pas faible par rapport à  $V(V)$ . La valeur assez élevée de  $V(R_{wv})^{-1}V_w R_{wv}^2 V(V_w)$  pour ces trois variables montre qu'il est important de tenir compte de la variance  $V(V_w)$  quand on étudie la stabilité de  $R_{wv}$ . La même remarque s'applique à l'effet de  $V(V_w)$  sur la stabilité de  $V_B$ , pour BPIK5.

6. DISCUSSION

Le présent document expose trois grandes idées. Tout d'abord, à cause du rôle que la variance intra-UPF estimée  $V_w$  joue dans le plan d'échantillonnage d'une enquête et dans l'analyse de ces données, il importe de tenir compte de l'erreur d'échantillonnage dans l'estimation de  $V_w$ . En deuxième lieu, les méthodes d'estimation habituelles qui s'articulent sur le plan d'échantillonnage débouchent sur des estimateurs relativement simples de la variance de  $V_w$  attribuable à l'échantillonnage. En général, ces mesures de la

stabilité doivent être interprétées avec prudence. Néanmoins, elles peuvent devenir des instruments de diagnostic utiles en permettant l'identification des variables pour lesquelles l'instabilité de  $V_w$  pose des difficultés particulières ou qui ont un effet prononcé sur la variance des valeurs associées comme  $V_B$  et  $R_{wv}$ . Troisièmement, l'application de ces méthodes à la U.S. Third National Health and Nutrition Examination Survey (NHANES III) et les simulations connexes révèle ce qui suit:

(i) pour des jeux de variables différents, les mesures de la stabilité  $d_{w0}$  observées sont cohérentes avec des conditions de stabilité sensiblement distinctes;

(ii) pour certaines variables, les estimateurs  $V_w$  sont considérablement moins stables que le laisserait supposer le dénombrement direct des unités secondaires d'échantillonnage;

(iii) pour quelques variables, la variance estimée de  $V_w$  contribue sensiblement à la variance estimée de la variance inter-UPF estimée  $V_{Bh}$  et au ratio des variances  $R_{wv}$ .

REMERCIEMENTS

Les auteurs aimeraient remercier Van Parsons, Cliff Johnson et d'autres statisticiens du NCHS pour leur avoir donné accès aux données de la NHANES III et avoir partagé avec eux une foule d'informations sur la phase I de cette enquête. Les auteurs remercient aussi Van Parsons et les deux examinateurs qui ont gardé l'anonymat pour leurs commentaires utiles sur les versions antérieures du document. La présente recherche a bénéficié en partie d'une aide du National Center for Health Statistics. Les points de vue exprimés dans ce document n'engagent que leurs auteurs et n'épousent pas nécessairement les politiques du National Center for Health Statistics.

BIBLIOGRAPHIE

FULLER, W.A. (1984). Application de la méthode du moindre carré et de techniques connexes aux plans de sondage complexe. *Techniques d'enquête*, 10, 107-130.

HANSEN, M.H., HURWITZ, W.N., et MADOW, W.G. (1953). *Sample Survey Methods and Theory, Volume I: Methods and Applications*. New York: John Wiley.

HERZOG, T.N., et SCHEUREN, F.J. (1976). Dallying with some CPS design effects for proportions. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 396-401.

JANG, D.S., et ELTINGE, J.L. (1996). Use of Within-PSU Variances and Errors-in-Variables Regression to Assess the Stability of a Standard Design-Based Variance Estimator. Manuscript inédit, Department of Statistics, Texas A&M University.

KORN, E.L., et GRAUBARD, B.G. (1995). Analysis of large health surveys: accounting for the sampling design. *Journal of the Royal Statistical Society, série A*, 158, 263-295.

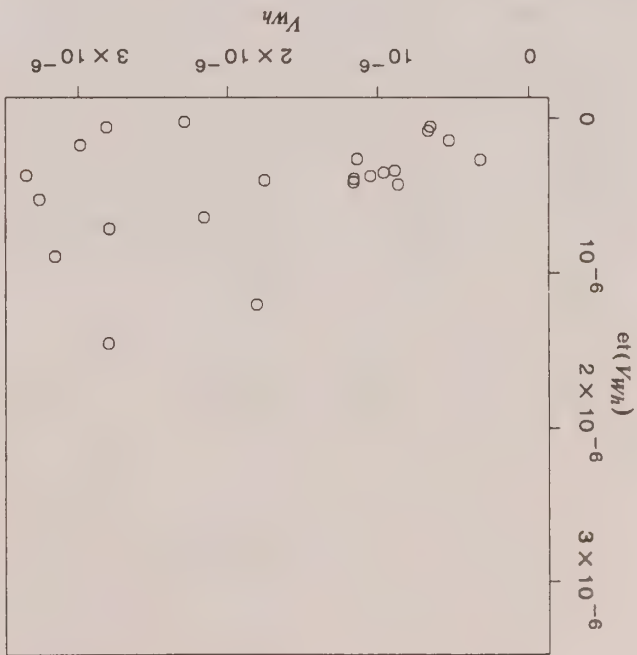


Figure 1. Diagramme de  $V(V_{wh})^{1/2}$  par rapport à  $V_{wh}$  pour HAE2

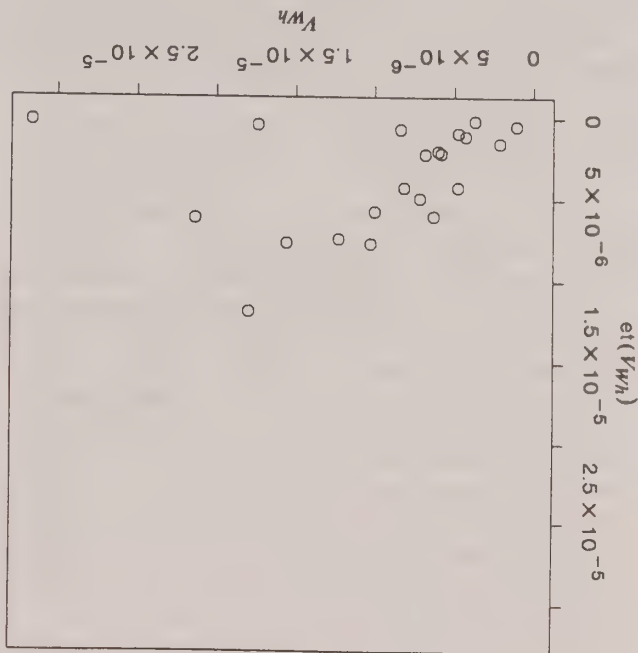


Figure 2. Diagramme de  $V(V_{wh})^{1/2}$  par rapport à  $V_{wh}$  pour log (concentration de plomb dans le sang)

liberté de  $d^{w_0} = 23.7$ . La variable log (concentration de plomb dans le sang) donne un peu plus de points près de la ligne de pente égale à un et d'ordonnée à l'origine égale à zéro, ce qui est conforme avec la valeur légèrement plus faible  $d^{w_0} = 10.5$ . Le diagramme de la concentration de plomb dans le sang présente une aberration: la plus forte valeur de  $V(V_{wh})^{1/2}$  correspond approximativement à la valeur de  $V_{wh}$ . On a examiné les termes  $V_{wh}$  et  $P_{hi}^{-2} \hat{O}_{2hi}$  de cette strate afin de déceler les tendances inhabituelles, c'est-à-dire les valeurs

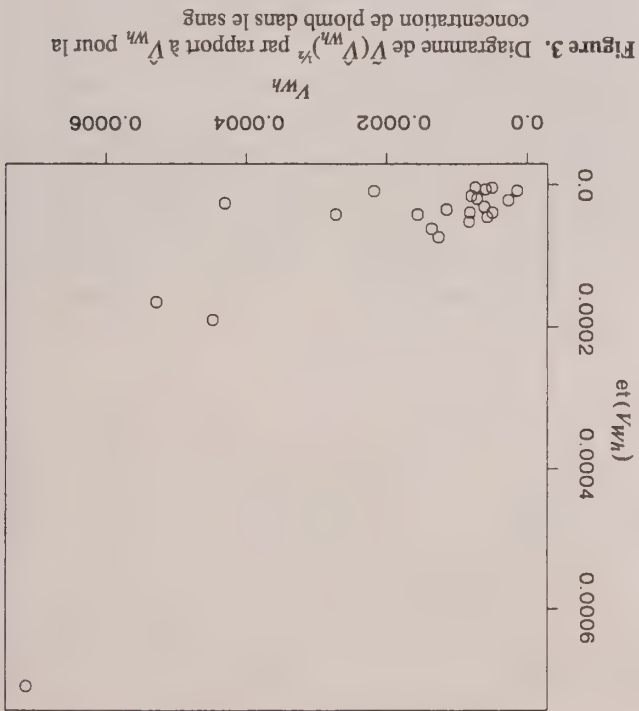


Figure 3. Diagramme de  $V(V_{wh})^{1/2}$  par rapport à  $V_{wh}$  pour la concentration de plomb dans le sang

### 5.3 Estimation de la variance inter-UPF et du ratio de la variance $R_{wv}$

Le tableau 5 donne les estimations  $V_B$  et  $R_{wv}$ , et l'erreur-type qui y est associée, pour les douze variables de la NHANES III. On s'intéressera en particulier à la colonne variance  $V(V_B)$  attribuable à la variance intra-UPF, et à la colonne  $V(R_{wv})^{-1} V(V_{wv})$ , la proportion correspondante de  $R_{wv}$ . La valeur relativement importante des proportions révèle que  $V(V_{wv})$  contribue sensiblement à  $V(V_B)$  et à  $R_{wv}$  pour les variables concernées. Remarquons que le rapport  $V(R_{wv})^{-1} V(V_{wv})$  est égal ou supérieur à 0.3 pour la concentration de plomb dans le sang, BPIK1 (pression artérielle systolique) et BPIK5 (pression artérielle diastolique). Les fortes proportions notées pour la concentration de plomb dans le sang et BPIK1 sont principalement attribuables à la valeur relativement importante

Le tableau 5 donne les estimations  $V_B$  et  $R_{wv}$ , et l'erreur-

raison de facteurs environnementaux.

Les fortes concentrations de plomb tendent à se regrouper en donner une distribution logarithmique à peu près normale et des E.-U., le dosage du plomb dans le sang a tendance à égard, on remarquera qu'avec une population comme celle et non du plan d'échantillonnage ou de la pondération. À cet sang pourrait donc résulter de quelques valeurs très élevées, inhabituelle observée pour la concentration de plomb dans le log (concentration de plomb dans le sang). La tendance valeur  $V(V_{wh})^{1/2}$  et  $V_{wh}$  aberrante pour d'autres variables, p. ex., plus forte valeur  $V_{wh}$ . Elle ne comporterait toutefois pas de  $P_{hi}^{-2} \hat{O}_{2hi}$  de l'UPF. Par ailleurs, cette strate présentait aussi la zéro tandis que la seconde était la plus élevée des termes Une des deux valeurs  $P_{hi}^{-2} \hat{O}_{2hi}$  associées était presque égale à extrêmes ou les poids excessifs au niveau de chaque élément.



Tableau 4  
Quantiles de  $d_{w0}$  obtenus par simulation

Cas	Moyenne	E.-T.	q.005	q.01	q.025	q.05	q.10	q.25	q.50	q.75	q.90	q.95	q.975	q.99	q.995
1	48.9	17.7	21.1	22.5	24.8	27.4	30.7	36.7	45.5	57.4	71.2	81.5	92.6	108.5	122.1
2	48.3	17.5	20.7	21.9	24.2	26.8	29.9	36.3	45.2	56.6	70.2	80.3	92.0	106.2	118.0
3	11.3	4.7	4.1	4.5	5.1	6.4	8.0	10.3	13.5	17.3	20.0	23.0	26.8	30.1	30.1
4	5.5	2.7	1.4	1.6	2.0	2.7	3.7	5.0	6.8	8.9	10.5	12.1	14.8	16.7	16.7
5	5.5	2.7	1.4	1.6	1.9	2.7	3.7	5.0	6.7	8.9	10.6	12.1	14.1	16.1	16.1
6	3.5	2.1	0.7	0.8	1.0	1.2	1.5	2.1	3.0	4.4	6.0	7.4	8.8	11.2	12.6

Pour le cas 1,  $d_{hi} = 22$  et  $R_{12} = 1$  sont utilisés. L'argumentation de la partie 3.3 révèle que les valeurs  $V_{wn}$  résultantes sont distribuées sous forme de multiples constants d'une variable aléatoire de distribution chi carré à  $d_{w0} = 44$  degrés de liberté. Pour ce cas, le choix de  $d_{hi} = 22$  pour la simulation débouche donc sur des quantiles de  $d_{w0}$  qui correspondent approximativement à ceux auxquels on s'attendrait avec le nombre moyen de 45,8 USE observé à la phase I de la NHANES III, dans les conditions décrites à la partie 3.4. Signalons que même dans les conditions idéales du cas 1,  $d_{w0}$  se caractérise par une variabilité relative passablement élevée. Comparons maintenant les valeurs  $d_{w0}$  du tableau 2 aux quantiles obtenus par simulation pour le cas 1. Les douze valeurs observées de  $d_{w0}$  sont inférieures à la valeur 24.8 du quantile 0.025 et dix des douze valeurs sont inférieures à la valeur 21.1 du quantile 0.005. Les valeurs  $d_{w0}$  relevées pour les variables de la NHANES III ne sont donc pas cohérentes avec la valeur nominale  $d_{w0} = 44$  obtenue dans les conditions idéales du cas 1.

### 5.2.3 Simulation dans d'autres conditions avec réduction de $d_{w0}$

En général, la distribution de  $d_{w0}$  peut s'écarter de celle qu'on observe dans le cas idéal (cas 1) pour les raisons suivantes: a) variabilité du nombre véritable  $n_{hi}$  d'USE; b) stabilité restreinte des estimations  $\hat{\sigma}_{2hi}^2$  au niveau des USE; et c) hétérogénéité des termes réels  $\hat{\sigma}_{2hi}^2$  au niveau de l'UPF. Les cas 2 à 6 portent sur les effets combinés de ces trois facteurs.

Le plan d'échantillonnage du cas 2 est identique à celui du cas 1, si ce n'est que  $d_{hi}$  est constitué de variables aléatoires ayant la même probabilité d'être sélectionnées dans l'échantillon avec remise des 44 valeurs  $n_{hi} - 1$  qui correspond aux 44 dénombrements  $n_{hi}$  des USE de l'ensemble de données original. Les quantiles de  $d_{w0}$  issus de la simulation sont semblables à ceux du cas 1.

Pour le cas 3,  $d_{hi} = 5$  et  $R_{12} = 1$ ; les valeurs  $V_{wn}$  résultantes sont distribuées comme des multiples constants des variables aléatoires de distribution chi carré à  $d_{w0} = 10$  degrés de liberté. Les quantiles du cas 3 obtenus par simulation sont légèrement plus cohérents avec les valeurs  $d_{w0}$  des données de la NHANES III. Ainsi, dix des douze variables ont une valeur  $d_{w0}$  égale ou supérieure à la valeur 6.4 du quantile 0.10. Deux variables (plomb et pression artérielle systolique) ont néanmoins une valeur  $d_{w0}$  inférieure à celle du quantile 0.005, obtenue par simulation pour le cas 3.

### 5.2.4 Diagramme de diagnostic

Au sens strictement numérique,  $d_{w0}$  dépend de l'ordre de grandeur de  $V(V_{wn})$  par rapport à  $2V_{wn}^2$ . Par conséquent, les diagrammes de diagnostic de  $V(V_{wn})^{1/2}$  par rapport à  $V_{wn}$  permettent d'identifier des tendances précises et les strates «problématiques» qui aboutissent à une valeur exagérément élevée ou faible de  $d_{w0}$ .

Les figures 1 à 3 présentent les diagrammes des variables HAF2 (hypertension), log (concentration de plomb dans le sang) et concentration de plomb dans le sang, respectivement. Chaque diagramme a un axe des abscisses et des ordonnées de la même échelle. La majorité des points du diagramme HAF2 se trouvent nettement sous une ligne de pente = 1 et d'ordonnée à l'origine = 0. De plus, les valeurs absolues élevées de  $V(V_{wn})^{1/2}$  restent sensiblement inférieures à la valeur de  $V_{wn}$  correspondante. Cette observation est cohérente avec le nombre relativement important de degrés de

Liberté. Les quantiles du cas 3 obtenus par simulation sont légèrement plus cohérents avec les valeurs  $d_{w0}$  des données de la NHANES III, dans les conditions décrites à la partie 3.4. Signalons que même dans les conditions idéales du cas 1,  $d_{w0}$  se caractérise par une variabilité relative passablement élevée. Comparons maintenant les valeurs  $d_{w0}$  du tableau 2 aux quantiles obtenus par simulation pour le cas 1. Les douze valeurs observées de  $d_{w0}$  sont inférieures à la valeur 24.8 du quantile 0.025 et dix des douze valeurs sont inférieures à la valeur 21.1 du quantile 0.005. Les valeurs  $d_{w0}$  relevées pour les variables de la NHANES III ne sont donc pas cohérentes avec la valeur nominale  $d_{w0} = 44$  obtenue dans les conditions idéales du cas 1.

Par ailleurs, les trois valeurs les plus élevées de  $d_{w0}$  (hypertension, cholestérolémie et pression artérielle diastolique) se retrouvent au-dessus de celle du quantile 0.995, dans chacun des cas 4 à 6. Compte tenu des résultats des cas 1 à 3 présentés plus haut, il semble donc que les douze valeurs  $d_{w0}$  observées soient cohérentes avec les conditions qui débouchent sur des valeurs réelles  $d_{w0}$  sensiblement différentes pour des variables distinctes.

Quant on les regroupe, les résultats de ces simulations donnent à penser que l'instabilité de  $V_{wn}$  pourrait être sensiblement plus grave que celle prévue d'après un simple dénombrement des USE de chaque strate, pour les douze variables de la NHANES III; en outre, il se pourrait que les véritables mesures de la stabilité  $d_{w0}$  fluctuent sensiblement d'une variable à l'autre.

Tableau 2  
Estimation de la variance et mesure de la stabilité des douze variables de la NHANES III

Variable	$V^w$	$V(V)$	$d^{w0}$	$d^{wf}$
HAE2	0.0000385	0.0000511	23.7	425.8
HAE7	0.0000821	0.000135	13.6	225.6
HAD1	0.00000956	0.00000749	8.8	160.6
HAR3	0.000122	0.000205	6.4	125.8
BMPHT	0.0223	0.0416	15.3	275.1
BMPWT	0.104	0.122	8.6	139.2
HDRESULT	0.0743	0.163	11.5	196.2
TCRESULT	0.590	0.860	21.2	353.9
LEAD	0.00388	0.00657	2.8	48.8
log(LEAD)	0.000211	0.000678	10.5	174.9
BPIK1	1.073	2.896	1.0	26.5
BPIK5	0.252	0.217	17.2	52.9

On a estimé le ratio pondéré  $\theta$  type de la valeur moyenne de chacune des douze variables énumérées au tableau 1. Les deux premières colonnes du tableau 2 donnent l'estimation correspondante de la variance  $V(\theta)$  et  $V^w$ . On s'est concentrablement intéressé à la stabilité des estimations  $V^{wn}$  dans le cadre d'une étude plus importante sur la variance intra-UPF  $V^{wn}$  rapportée dans Jang et Eitinge (1996). Puisque  $n_h = 2$  dans chaque strate, le raisonnement de la partie 3.2 révèle qu'on ne peut examiner les termes  $d^{wn}$  séparément. C'est pourquoi on étudiera la mesure groupée  $d^{w0}$  de la stabilité de  $V^{wn}$  à la partie 5.2 et proposera des tests de simulation et des diagrammes de diagnostic connexes.

Par ailleurs, on était intéressé à déterminer l'apport des variances de  $V^{wn}$  à la variance des quantités groupées  $V^B$  et  $R^{wv}$ . La partie 5.3 explore cet aspect.

5.2 Estimation de la variance intra-UPF et mesures de la stabilité connexes

5.2.1 Comparaison des variables

Les deux dernières colonnes du tableau 2 donnent le nombre estimé de degrés de liberté  $d^{w0}$  et  $d^{wf}$  des douze variables de la NHANES III. Notons que la stabilité  $d^{w0}$  de la strate est relativement faible comparativement à la moyenne de 45.8 USF par strate. Toutes les variables présentent notamment une valeur  $d^{w0}$  inférieure à 24, et cette valeur est inférieure à 10 dans cinq cas (HAD1, HAR3, BMPWT, LEAD et BPIK1). Face à l'intérêt que suscitent les valeurs  $d^{w0}$  mentionnées ci-dessus, on se pose deux grandes questions:

(1) Les valeurs  $d^{w0}$  observées sont-elles cohérentes avec le nombre nominal de degrés de liberté  $d^{w0}$  auquel on s'attendrait avec le dénombrement direct  $n_{h1} + n_{h2} - 2$  des USF?

(2) Inversement, les valeurs  $d^{w0}$  observées sont-elles cohérentes avec les conditions de distribution qui débouchent sur des valeurs considérablement plus faible de  $d^{w0}$ ?

Les tests types s'inspirant de la théorie des grands échantillons pour (1) et (2) auraient du s'appuyer sur les moments du huitième échantillon. Il n'aurait pas été conseillé de s'en servir ici, étant donné les valeurs relativement faibles  $L = 22$  et  $n_h = 2$ . C'est pourquoi on a procédé au test de simulation qui suit.

5.2.2 Interprétation des mesures de stabilité par simulation

La simulation portait sur six cas caractérisés par deux termes de valeur différente. Le premier terme, noté  $d_{hi}$ , représente le nombre de degrés de liberté associé à l'estimateur de la variance  $\hat{\sigma}_{2hi}^2$  de l'UPF  $(h, i)$ . Le second, noté  $R_{12}$ , est le rapport des expressions  $P_{hi}^{-2} \hat{\sigma}_{2hi}^2$  pour la première et la deuxième UPF de l'échantillon de la strate  $h$ .

Pour chacun des six cas présentés plus bas, on a obtenu les variables indépendantes pseudoaléatoires  $g_{hi}$  d'une distribution chi carré à  $d_{hi}$  degrés de liberté, où  $h = 1, 2, \dots, 22$  et  $i = 1, 2$ . On a ensuite calculé les variables rééchantonnées  $V^{whi} = d_{hi}^{-1} V^{whi} g_{hi}$ , où  $V^{whi}$  représente une variable aléatoire dont la probabilité de prendre soit la valeur un, soit la valeur  $R_{12}$ , est égale à la dernière. Les variables aléatoires  $g_{hi}$  et  $V^{whi}$  sont indépendantes l'une de l'autre. Enfin, on a calculé les sommes  $V^{wn} = V^{wh1} + V^{wh2}$  et les valeurs connexes  $V(V^{wn})$ ,  $V^w$  et  $d^{w0}$ . Cet exercice a été répété 10,000 fois.

Le tableau 3 présente les valeurs de  $d_{hi}$  et  $R_{12}$  pour les six cas qui nous intéressent tandis que la moyenne, l'écart-type et les quantités de  $d^{w0}$  de la simulation apparaissent au tableau 4. Lorsqu'on analyse les résultats, on remarque que le caractère aléatoire de  $g_{hi}$  correspond à l'erreur d'estimation de  $\hat{\sigma}_{2hi}^2$  attribuable au sous-échantillonnage de l'USF et des niveaux inférieurs; le caractère aléatoire de  $V^{whi}$  reflète pour sa part la variabilité de  $P_{hi}^{-2} \hat{\sigma}_{2hi}^2$  induite par l'échantillonnage des UPF d'une strate.

Cas couverts par les quantiles de la simulation			
Cas	$d$	$R_{12}$	
1	22		1
2		Distr. Obs.	1
3	5		1
4	22		9
5		Distr. Obs.	9
6	5		9



Pour le reste de la partie 4.2, nous supposons aussi que  $\text{Cov}(V_{wh}^h, V_h^h) = 0$ . Les arguments ordinaires conditionnels au moment révèlent que l'hypothèse tient si les termes  $p_{hi}^{2hi} o_{2hi}^{2hi}$  sont égaux pour une strate donnée et si les estimations  $x_{hi}^{hi}$  au niveau de l'USF connaissent une distribution normale, sous réserve de  $(h, i, j)$ , de sorte que  $\delta_{2hi}^{2hi}$  est indépendant de  $X_{hi}^{hi}$  sous certaines conditions.

4.2.2 Mesures de la stabilité

Compte tenu des conditions énoncées à la partie 4.2.1,  $V(V_B^h)$  et  $V(V_B^h)$  ont pour estimateurs non biaisés

$$V(V_B^h) = (n_h + 1)^{-1} 2V_h^2 + V(V_{wh}^h) \quad (4.1)$$

et  $V(V_B^h) = \sum_{h=1}^L V(V_B^h)$ , où  $V(V_{wh}^h)$  est défini dans l'expression (2.2). Toujours en vertu des mêmes conditions, les arguments ordinaires relatifs à l'estimation du ratio débouchent sur l'estimateur de la variance

$$V(R_{wv}^h) = V^{-2} \sum_{h=1}^L \left\{ (n_h + 1)^{-1} 2V_h^2 + R_{wv}^h V(V_{wh}^h) \right\}. \quad (4.2)$$

4.3 Autres mesures de la stabilité par groupement des strates

Dans certains cas, les hypothèses énoncées à la partie 4.2.1 peuvent soulever des difficultés. Ainsi, les estimateurs  $x_{hi}^{hi}$  de l'USF peuvent présenter une distribution nettement atypique avec certains plans d'échantillonnage et variables, si bien que l'hypothèse  $\text{Cov}(V_{wh}^h, V_h^h) = 0$  doit être rejetée. Quand cela se produit, on peut envisager le groupement des strates afin d'obtenir d'autres estimateurs de  $V(V_B^h)$  et de  $V(R_{wv}^h)$ . Plus exactement, en séparant la série de  $L$  strates en  $G$  groupes pré-établis, le groupe  $S_g$ ,  $g = 1, \dots, G$  contenant  $L_g$  strates, on constate que

$$(V(V), V_w, V_B) = \sum_{g=1}^G \sum_{h \in S_g} (V_h, V_{wh}, V_{Bh}).$$

Les méthodes habituelles utilisées pour grouper les strates (lire Wolter 1985, partie 2.5) permettent d'obtenir l'autre estimateur de variance

$$V^*(V_B^h) = \sum_{g=1}^G (L_g - 1)^{-1} L_g \sum_{h \in S_g} D_{gh}^2,$$

où  $D_{gh} = V_{Bh}^h - L_g^{-1} \sum_{j \in S_g} V_{Bj}^h$ . L'estimateur de la variance de  $R_{wv}^h$  obtenu par groupement des strates correspond à

$$V^*(R_{wv}^h) = (V_{wv}^h)^{-2} \sum_{g=1}^G (L_g - 1)^{-1} L_g \sum_{h \in S_g} C_{gh}^2,$$

où  $C_{gh} = (V_h^h - R_{wv}^h V_{wh}^h) - L_g^{-1} \sum_{j \in S_g} (V_j^h - R_{wv}^h V_{wj}^h)$ . En règle générale, les estimateurs de variance qui dérivent du groupement des strates doivent être jugés avec prudence; lire par exemple Rust et Kalton (1985), Wolter (1985, partie 2.5) et les ouvrages cités dans la bibliographie.

Les estimateurs de variance obtenus par groupement des

strates sont habituellement conservateurs. De plus, avec un nombre  $L$  moyen, les estimateurs  $V^*(V_B^h)$  et  $V^*(R_{wv}^h)$  peuvent eux-mêmes être caractérisés par une stabilité restreinte.

5. APPLICATION À LA U.S. THIRD NATIONAL HEALTH AND NUTRITION EXAMINATION SURVEY

5.1 Plan d'échantillonnage et méthodes d'estimation

Les méthodes présentées aux parties 2 à 4 ont été appliquées aux données de la phase I de la U.S. Third National Health and Nutrition Examination Survey (NHANES III). On trouvera dans National Center for Health Statistics (1996) une description générale de l'enquête avec les particularités de la phase I (données recueillies entre 1988 et 1991). Aux fins de la discussion, trois aspects nous intéressent plus spécialement. Tout d'abord, les estimateurs de la variance ont été construits selon un plan d'échantillonnage groupé comprenant  $L = 22$  strates (groupes importants de comtés), avec sélection de deux unités primaires d'échantillonnage (habituellement des comtés) par strate. Deuxièmement, chaque UPF choisie comptait un nombre relativement important d'USF (généralement, des groupes de quartiers ou des régions rurales analogues). Le nombre d'USF sélectionnées dans chaque strate variait de 30 à 63, avec une moyenne de 45,8. Un troisième échantillonnage dans chaque USF a permis d'obtenir les éléments de l'enquête (civils américains non institutionnalisés). Chaque personne choisie a été prise de répondre à un questionnaire sur la santé et de subir un examen médical poussé. Douze des variables résultantes apparaissent au tableau 1.

Tableau 1  
Douze variables de la NHANES III

Nom de la variable	Description
HAE2	A appris d'un professionnel de la santé qu'il souffrait d'hypertension (variable indicatrice)
HAE7	A appris d'un professionnel de la santé que sa cholestérolémie était élevée (variable indicatrice)
HAD1	A appris d'un professionnel de la santé qu'il souffrait du diabète (variable indicatrice)
HAR3	Fumez-vous?
BMPHT	Hauteur
BMPWT	Poids
HDRESULT	Cholestérol HDL
TCRESULT	Cholestérolémie
LEAD	Concentration de plomb dans le sang,
log(LEAD)	microgrammes par décilitre
BPIK1	plomb dans le sang
BPIK5	Pression artérielle systolique
	Pression artérielle diastolique

Étant donné un estimateur non biaisé  $\hat{V}(V)$  de la variance de  $V$ , il est possible de calculer l'estimateur «degrés de liberté»  $\hat{d}$  en résolvant l'équation d'estimation non biaisée

$$(3.1) \quad 2\{\hat{V}^2 - \hat{V}(\hat{V})\} - \hat{V}(\hat{V})\hat{d} = 0,$$

c.-à-d.,  $\hat{d} = \{\hat{V}(\hat{V})\}^{-1} 2\hat{V}^2 - 2$ . En supposant une régularité moyenne, la probabilité de  $\hat{d}^{-1}\hat{d}$  converge vers un, pourvu que la probabilité de  $\{\hat{V}(\hat{V})\}^{-1}\hat{V}(\hat{V})$  et  $\{E(\hat{V})\}^{-1}\hat{V}$  en fasse autant.

### 3.2 Application de la méthode des degrés de liberté aux estimateurs de la variance intra-UPF groupée et de la variance de la strate

Les principes généraux sur les degrés de liberté peuvent s'appliquer aux estimateurs de la variance intra-UPF  $V_{wh}$  et  $V_{wv}$  développés à la partie 2. Examinons d'abord le cas où la stabilité des estimateurs individuels  $V_{wh}$  de la strate suscitent un intérêt intrinsèque. La mesure «degrés de liberté» connexe est  $d_{wh} = \{V(\hat{V}_{wh})\}^{-1} 2\hat{V}_{wh}^2$ . Avec les plans d'échantillonnage caractérisés par un  $n_h$  élevé, on pourrait se servir de (3.1) afin de calculer les estimateurs  $\hat{d}_{wh} = \{\hat{V}(\hat{V}_{wh})\}^{-1} 2\hat{V}_{wh}^2 - 2$  séparément pour chaque strate. Avec un faible  $n_h$  (à savoir, quand  $n_h = 2$  pour chaque strate), l'estimateur  $\hat{d}_{wh}$  peut être en soi très instable. C'est pourquoi, il vaut la peine d'examiner l'autre estimateur combiné

$$\hat{d}_{w0} = \left\{ \sum_L \hat{V}(\hat{V}_{wh}) \right\}^{-1} 2 \sum_L \hat{V}_{wh}^2 - 2,$$

en partant de l'hypothèse que tous les  $d_{wh}$  ont la même valeur  $d_{w0}$ . Examinons maintenant l'estimateur de variance intra-UPF groupée  $V_w$  de la partie 2.3. Le nombre de «degrés de liberté» résultant est  $d_{wv} = \{\sum_{h=1}^L \hat{V}(\hat{V}_{wh})\}^{-1} 2\hat{V}_w^2$  et l'expression (3.1) débouche sur l'estimateur

$$\hat{d}_{wv} = \left\{ \sum_L \hat{V}(\hat{V}_{wh}) \right\}^{-1} 2\hat{V}_w^2 - 2.$$

### 3.3 Comparaison de $d_{w0}$ et $d_{wv}$ au dénombrement direct des UPF

Pour utiliser  $\hat{d}_{w0}$  et  $\hat{d}_{wv}$  comme mesure de la stabilité, supposons les conditions idéales qui suivent. Pour chaque  $h$ , le nombre  $n_h$  d'UPF correspond à la valeur commune  $n_1$  par exemple; pour les valeurs  $h$  et  $i$ , le nombre d'UPF  $n_{hi}$  donne la valeur commune  $n_{11}$ . Enfin, supposons que les termes  $p_{hi} \sigma_{2hi}^{-2}$  sont constants pour chaque strate et que, sous réserve de  $(h,i)$ , chaque valeur  $\sigma_{2hi}^{-2}(n_{11} - 1) \phi_{2hi}^2$  soit distribuée de la même façon qu'une variable aléatoire de distribution chi carré à  $n_{11} - 1$  degrés de liberté. Les arguments ordinaires montrent alors que  $d_{w0} = n_1(n_{11} - 1)$ . Si les hypothèses qui précèdent trouvent à peu près confirmation et si le produit  $n_1(n_{11} - 1)$  est assez important (supérieur à 40, par exemple), l'utilisateur des données pourrait être enclin à considérer  $\hat{V}_{wh}$  comme étant relativement stable, ou estimer que les erreurs  $V_{wh} - V_{wh}$  sont négligeables, ce qui revient au même. Ce

## 4. COMPARAISON DE LA VARIANCE INTRA-UPF ET DE LA VARIANCE GLOBALE DE LA STRATE

### 4.1 Estimateurs de la variance inter-UPF et des ratios connexes des variances

La partie 1 mentionnait quelques applications reposant sur l'ordre de grandeur de  $V_w$  par rapport à  $V_h$ . Les particularités des comparaisons de l'ordre de grandeur varient avec l'application, mais en général, on s'intéresse surtout à la différence ou au ratio. Ainsi, on se rappellera que  $V_{Bh} = V_h - V_{wh}$  définit la variance inter-UPF globale  $V_B = \sum_{h=1}^L V_{Bh}$ . Par ailleurs,  $V_{Bh}$  et  $V_B$  ont pour estimateurs non biaisés  $\hat{V}_{Bh} = \hat{V}_h - \hat{V}_{wh}$  et  $\hat{V}_B = \sum_{h=1}^L \hat{V}_{Bh}$ , respectivement.

De même, soit le ratio  $R_{wv} = V^{-1}V(\hat{Y})$ , ordre de grandeur de la variance globale  $V(\hat{Y})$  par rapport à l'apport intra-UPF de  $V_w$ .  $R_{wv} = \hat{V}^{-1}\hat{V}(\hat{Y})$  est un estimateur direct de  $R_{wv}$ . Soulignons que si  $V^{-1}V_h = R_{wv}$  pour toutes les valeurs de  $h$ ,  $R_{wv}$  pourrait également être considéré comme un estimateur groupé du ratio commun de la strate.

Avec  $\hat{V}_B$  et  $R_{wv}$ , l'évaluation de la stabilité fait intervenir la variance de  $\hat{V}_B$  et la covariance de  $\hat{V}_{wh}$  avec  $\hat{V}_B$ . L'estimation de ces moments pose parfois un problème dans les enquêtes où on ne prélève qu'un petit nombre d'UPF dans chaque strate. Nous envisageons deux approches en guise de solution. La partie 4.2 impose des restrictions moyennes à la structure du moment de  $(\hat{V}_{wh}, \hat{V}_B)$  pour parvenir à l'estimateur  $V(\hat{V}_B)$  et aux valeurs connexes. La partie 4.3 recourt au groupement des strates pour établir d'autres mesures de la stabilité.

### 4.2 Mesure de la stabilité d'après $\hat{V}(\hat{V}_{wh})$ et les conditions du moment

4.2.1 Conditions du moment

Avec l'application de restrictions moyennes au moment, il est possible d'estimer la variance de  $\hat{V}_B$  directement à partir de  $\hat{V}_B$ . Plus précisément, supposons que la variance de  $\hat{V}_B$  soit égale à  $(n_h - 1)^{-1} 2\hat{V}_{wh}^2$ ; une telle supposition serait valable, par exemple, avec l'hypothèse type que  $V_h^{-1}(n_h - 1)V_h$  a la distribution d'une variable aléatoire de distribution chi carré à  $n_h - 1$  degrés de liberté. Comme on l'a fait aux parties 2 et 3, on présume aussi que  $\hat{V}_h$  n'est pas biaisé pour  $V_h$ . Les arguments ordinaires relatifs au moment montrent que  $(n_h + 1)^{-1} 2\hat{V}_{wh}^2$  est un estimateur non biaisé de la variance de  $\hat{V}_B$ .



## 2.3 Variance de $V^{w_h}$

Une modification directe des arguments types conditionnels au moment révèle que la variance de  $V^{w_h}$  est  $\gamma_{Bh} + \gamma_{wh}$ , où

$$\gamma_{Bh} = V(n_h^{-2} \sum_{h=1}^I p_{hi}^{-2} \sigma_{2hi}^2)$$

$$\gamma_{wh} = n_h^{-3} \sum_{h=1}^I p_{hi}^{-3} V(\sigma_{2hi}^2 | h, i).$$

et

La variance de  $V^{w_h}$  repose donc en soi sur la somme des variances inter-UPF et intra-UPF, tandis que l'ordre de

grandeur relatif de  $\gamma_{Bh}$  et  $\gamma_{wh}$  dérive d'un compromis entre  $\sigma_{2hi}^2$  et  $n_{hi}$ . Par exemple, en supposant la régularité, les termes  $V(\sigma_{2hi}^2 | h, i)$  sont presque inversement proportionnels à  $n_{hi}$ . Si  $n_{hi}$  est élevé partout dans la strate  $h$ ,  $\gamma_{wh}$  pourrait être relativement faible. Par ailleurs, si les termes  $p_{hi}^{-2} \sigma_{2hi}^2$  sont à peu près constants dans une strate donnée,  $\gamma_{Bh}$  pourrait être assez faible lui aussi. À l'inverse, une nette hétérogénéité de  $p_{hi}^{-2} \sigma_{2hi}^2$  pourrait accroître  $\gamma_{Bh}$ , et  $V(V^{w_h})$  par la même occasion.

Remarquons par ailleurs qu'en vertu des conditions établies par le plan d'échantillonnage,  $V^{w_h}$  représente la moyenne des termes  $n_h^{-1} p_{hi}^{-2} \sigma_{2hi}^2$  indépendant, mais à distribution identique, de l'échantillon. Un estimateur non biaisé de la variance de  $V^{w_h}$  est

$$V(V^{w_h}) = n_h^{-1} (n_h - 1)^{-1} \sum_{h=1}^I (n_h^{-1} p_{hi}^{-2} \sigma_{2hi}^2 - V^{w_h})^2. \quad (2.2)$$

Quelques applications portent plus sur l'ensemble de la population que sur les strates individuelles. Dans un tel cas, la contribution de la variance «intra-UPF» qui nous intéresse correspond à la somme des variances intra-UPF,  $V^w = \sum_{h=1}^I V^{w_h}$ . Compte tenu des conditions qui précèdent,  $V^w$  serait un estimateur non biaisé de  $V^w$ . D'autre part, puisque l'échantillonnage et le sous-échantillonnage sont indépendants d'une strate à l'autre,  $V(V^w) = \sum_{h=1}^I (\gamma_{Bh} + \gamma_{wh})$  et  $V(V^w)$  aurait pour estimateur non biaisé

$$V(V^w) = \sum_{h=1}^I V(V^{w_h}).$$

Enfin, soulignons que le développement qui précède repose sur l'hypothèse d'un échantillon avec remise au niveau tant de l'unité primaire que de l'unité secondaire. Deux applications du résultat (2.4.16) dans Wolter (1985, p. 86) montrent qu'avec de légères conditions, que respectent de nombreux plans d'échantillonnage avec remise mais pas tous, la valeur  $V^{w_h}$  n'est pas biaisée ou reste conservatrice pour la variance intra-UPF véritable. Il en va autant pour  $V(V^{w_h})$ , pour la variance véritable de  $V^{w_h}$ . Les auteurs sont en mesure de fournir une analyse et une preuve techniques formelles de ce résultat.

## 2.4 Interprétation équilibrée des mesures de la stabilité

Dans le reste du document,  $V(V^{w_h})$  et les valeurs connexes servent à évaluer la stabilité des estimateurs des composantes de la variance. Lorsqu'on se sert des résultats indiqués, il vaut la peine de se rappeler que les mesures de la stabilité des estimateurs de la variance qui dérivent des données doivent avec raison être jugées avec une certaine prudence, car elles dépendent des moments du quatrième échantillon, donc sont en soi sujettes à une variabilité considérable de l'échantillonnage. On lira à ce sujet Fuller (1984, p. 111). L'aver-tissement vaut aussi pour l'estimateur  $V(V^{w_h})$  proposé et les statistiques connexes présentées plus loin aux parties 3 et 4. Il ne faudrait toutefois pas pêcher par excès de prudence en évitant d'évaluer la stabilité des estimateurs d'après les données. L'estimateur  $V(V^{w_h})$  et les mesures connexes présentées aux parties 3 et 4 sont relativement simples à calculer et peuvent servir d'instrument de diagnostic pour identifier les variables suivantes:

- (a) celles pour lesquelles l'instabilité de  $V^{w_h}$  pose un problème particulier;
  - (b) celles pour lesquelles l'instabilité de  $V^{w_h}$  agit sensiblement sur la précision des estimateurs de l'ordre de grandeur relatif de la variance inter-UPF et intra-UPF.
- Par conséquent, en interprétant les valeurs spécifiques de  $V(V^{w_h})$  et les mesures connexes de la stabilité, on devrait parvenir à un juste compromis entre l'avertissement général qui précède et l'utilité potentielle des résultats sur le plan du diagnostic.

## 3. DEUX MESURES DE LA STABILITÉ DES ESTIMATEURS DE LA VARIANCE INTRA-UPF

### 3.1 Diagnostic de la stabilité de l'estimateur de la variance selon les degrés de liberté

Certains analystes préfèrent exprimer la stabilité de l'estimateur de la variance au moyen des «degrés de liberté» associés à l'approximation de Satterthwaite (1941, 1946). En guise d'introduction à cette méthode, soit un estimateur général de la variance  $V$  et soit  $\{E(V)\}^{-1} dV$ , dont les deux premiers moments sont identiques à ceux d'une variable aléatoire de distribution chi carré à  $d$  degrés de liberté, où  $d$  représente la solution de l'équation

$$2\{E(V)\}^{-2} - V(V)d = 0.$$

Si la distribution de  $\{E(V)\}^{-1} dV$  approche effectivement une distribution chi carré,  $d$  pourrait assez littéralement être considéré comme le nombre de «degrés de liberté». Dans les autres cas,  $d$  correspondrait au double de la valeur inverse du coefficient de variation de  $V$ , au carré. Dans l'un ou l'autre cas,  $d$  présente de l'intérêt, car le terme est indépendant de l'échelle et on peut le relier assez directement à la notion de «taille efficace de l'échantillon» quand on s'efforce d'évaluer la performance de l'estimateur de la variance. La partie 3.3 présente des commentaires pertinents sur ces deux cas particuliers.

La partie 5 reprend les idées principales des parties 2 à 4 en les appliquant à l'estimation de la variance de la troisième enquête nationale sur la santé et la nutrition des E.-U. On recourt aussi à une technique de simulation pour établir la cohérence des observations avec les hypothèses types concernant la stabilité de l'estimateur de la variance, à la partie 5. Les résultats de la partie 5 donnent à penser que la stabilité réelle des estimateurs de la variance intra-UPF est sensiblement plus faible que le laisse croire un simple dénombrement des unités secondaires de chaque UPF. Les mêmes résultats révèlent que les propriétés de stabilité des estimateurs de la variance intra-UPF et des valeurs connexes changent sensiblement d'une variable à l'autre, dans la même enquête. Enfin, la partie 6 présente des commentaires supplémentaires sur les méthodes et les résultats empiriques mentionnés dans le document.

## 2. ESTIMATEUR DE LA VARIANCE INTRA-UPF ET DE LA STRATE GLOBALE

### 2.1 Système de notation général

En principe, on pourrait étudier les composantes de la variance intra-UPF et inter-UPF grâce à des méthodes reposant sur le plan d'échantillonnage ou à un modèle. Nous avons pour notre part retenu l'approche du plan d'échantillonnage. Cette dernière est cohérente avec celle mentionnée par plusieurs auteurs, notamment Wolter (1985, p. 40-41, 47). L'approche du plan d'échantillonnage s'avère particulièrement utile pour mettre en relief quelques points forts et quelques lacunes des méthodes suggérées pour évaluer la stabilité. Ainsi, à la partie 2.3, cette approche nous donnera une idée des particularités du plan qui peuvent agir sur la stabilité de l'estimateur de la variance. À la partie 4, la même approche nous aidera à éclaircir la mesure dans laquelle certaines restrictions du moment justifient une série de mesures de la stabilité.

Reprenant la notation et les idées de Wolter (1985, p. 43-47), soit un plan d'échantillonnage stratifié à plusieurs degrés comportant  $L$  strates et  $N_h$  unités primaires d'échantillonnage (UPF) dans la strate  $h = 1, 2, \dots, L$ . Nous sélectionnons  $n_h$  UPF d'un échantillon avec remise, avec probabilité de tirage au sort de  $p_{hi}$ . Pour chaque UPF  $(h, i)$  sélectionnée, on prélèvera  $n_{hi}$  unités secondaires d'échantillonnage (USF) d'un échantillon avec remise, avec probabilité de  $p_{hij}$  d'être tiré au sort. Le sous-échantillonnage se poursuivra à l'intérieur de l'USF sélectionnée jusqu'à ce qu'on obtienne  $n_{hij}$  sujets pour l'interview ou l'examen. Les méthodes d'évaluation de la stabilité élaborées ici sont principalement destinées aux plans d'échantillonnage caractérisés par un  $L$  moyen ou élevé, un  $n_h$  relativement faible (à savoir,  $n_h = 2$ ) et un  $n_{hi}$  assez important. Les plans présentant ces caractéristiques correspondent à ceux utilisés dans le cadre des enquêtes-ménages par interview de grande envergure, notamment l'enquête sur la santé dont il est question à la partie 4.

### 2.2 Variance intra-UPF et inter-UPF

Notre estimation portera sur une population de  $X = \sum_{h=1}^L X_h$ , où  $X_h = \sum_{i=1}^{N_h} X_{hi}$ ,  $X_{hi} = \sum_{k=1}^{N_{hi}} X_{hik}$ ,  $X_{hik}$  représente l'item de l'enquête pour l'élément  $k$  de l'USF  $j$  de l'UPF  $i$  de la strate  $h$ ,  $N_{hi}$  correspond au nombre d'USF de l'UPF  $(h, i)$ , et où  $N_{hij}$  est égal au nombre d'éléments de l'USF  $(h, i, j)$ . L'extension du chiffre de la population en fonction non linéaire est assez simple; nous y reviendrons à la partie 5 sur les applications. Un estimateur type de  $X$  reposant sur le plan de l'échantillonnage est  $\hat{X} = \sum_{h=1}^L \hat{X}_h$ , où

$$\hat{X}_h = \sum_{i=1}^{N_h} \hat{X}_{hi} = \sum_{i=1}^{N_h} p_{hi}^{-1} \sum_{j=1}^{N_{hi}} \sum_{k=1}^{N_{hij}} w_{hijk} y_{hijk}, \quad (2.1)$$

où  $\hat{X}_{hi} = n_{hi}^{-1} \sum_{j=1}^{N_{hi}} z_{hij}$  et  $z_{hij} = n_{hi} n_{hi} p_{hi}^{-1} \sum_{k=1}^{N_{hij}} w_{hijk} y_{hijk}$ . Il vaudra la peine de récrire l'expression (2.1) comme suit

Un estimateur non biaisé ordinaire de la variance globale de la strate  $V_h$  est

$$V(\hat{X}_h) = n_h^{-1} \sum_{i=1}^{N_h} (p_{hi}^{-1} \hat{X}_{hi} - \hat{X}_h)^2,$$

l'estimateur correspondant de  $V(\hat{X}) = \sum_{h=1}^L V(\hat{X}_h)$  est l'estimateur correspondant de  $V(\hat{X}) = \sum_{h=1}^L V(\hat{X}_h)$ . Passons maintenant à l'estimation de la variance intra-UPF  $V_{wh}$ . Puisque  $\hat{X}_{hi}$  représente la moyenne de l'échantillon pour les termes indépendants, à distribution identique,  $z_{hij}$ , l'argumentation habituelle nous apprend que pour l'UPF  $(h, i)$ ,  $\sigma_{2hi}^2$  est un estimateur non biaisé de  $\sigma_{2hi}^2 = n_{hi}^{-1} (n_{hi} - 1)^{-1} \sum_{j=1}^{N_{hi}} (z_{hij} - \bar{X}_{hi})^2$ . Par conséquent, l'estimateur non biaisé de  $V_{wh}$  est

$$V_{wh} = \sum_{i=1}^{N_h} n_h^{-2} p_{hi}^{-2} \sigma_{2hi}^2 = \sum_{i=1}^{N_h} n_{hi}^{-1} (n_{hi} - 1)^{-1} \sum_{j=1}^{N_{hi}} (x_{hij} - \bar{x}_{hi})^2,$$

où  $x_{hij} = n_{hi} \sum_{k=1}^{N_{hij}} w_{hijk} y_{hijk}$  et  $\bar{x}_{hi} = n_{hi}^{-1} \sum_{j=1}^{N_{hi}} x_{hij}$ . Notons que la dernière expression de  $V_{wh}$  repose uniquement sur la taille de l'échantillon, les observations  $y_{hijk}$  et les poids ordinaires  $w_{hijk}$ .



Mesures de la stabilité des estimateurs des composantes de la variance dans un plan d'échantillonnage stratifié à plusieurs degrés

J.L. ELTINGE et D.S. JANG<sup>1</sup>

## RÉSUMÉ

Les travaux sur les enquêtes par échantillonnage exigent souvent qu'on recoure aux estimateurs des composantes de la variance associées à l'échantillonnage, à l'intérieur des unités primaires d'échantillonnage et entre celles-ci. Dans ce genre de travail, il peut s'avérer important d'avoir une idée de la stabilité des estimateurs des composantes de la variance, bref de savoir si ces estimateurs présentent une variance relativement faible. Nous examinerons ici plusieurs façons de mesurer la stabilité des estimateurs des composantes de la variance reposant sur le plan d'échantillonnage et des quantités connexes, d'après les données. Dans le développement, on mettra en relief les méthodes applicables aux enquêtes caractérisées par un nombre moyen ou important de strates et un petit nombre d'unités primaires d'échantillonnage par strate. Nous attirons principalement l'attention sur la variance intrinsèque d'un estimateur de la variance intra-UPF et sur deux termes connexes se rapportant aux degrés de liberté. Une méthode de simulation permet d'établir si la stabilité observée est cohérente avec les hypothèses types sur la stabilité de l'estimateur de la variance. Nous présentons aussi deux séries de mesures de stabilité pour les estimateurs des composantes de la variance inter-UPF reposant sur le plan d'échantillonnage et le ratio de la variance globale avec la variance intra-UPF. Les méthodes proposées sont appliquées aux données venant des interviews et des examens de la U.S. Third National Health and Nutrition Examination Survey (NHANES III). Les résultats montrent que les propriétés de la stabilité véritable peuvent changer sensiblement d'une variable à l'autre. Par ailleurs, pour certaines variables, les estimateurs de la variance intra-UPF semblent considérablement moins stables qu'on aurait pu s'y attendre consécutivement à un simple dénombrement des unités secondaires de chaque strate.

MOTS CLÉS: Variance inter-UPE; plan d'échantillonnage complexe; degrés de liberté; diagnostic; analyse selon le plan d'échantillonnage; approximation de Satterthwaite; groupement des strates; U.S. Third National Health and Nutrition Examination Survey (NHANES III); variance intra-UPE

## I. INTRODUCTION

Avec les enquêtes par échantillonnage, il vaut souvent la peine de bien estimer les composantes de la variance qui résulte de l'échantillonnage, à l'intérieur des unités primaires d'échantillonnage (UP) et entre celles-ci. Ainsi, l'ordre de grandeur de la variance intra-UP estime par rapport à la variance inter-UP pourrait exercer une influence sur la division de l'échantillon et les questions connexes relative au plan d'échantillonnage (lire Hansen et coll., 1953, chapitre 7). Des propriétés analogues concernant l'ordre de grandeur modifient le biais de certains estimateurs de la variance obtenus lorsqu'on simplifie les hypothèses du plan d'échantillonnage (lire Korn et Graubard 1995, p. 278-279, 287; et Wolter 1985, p. 44-46). Par ailleurs, certains analystes expriment un intérêt général pour l'identification des enquêtes et des variables où la composante inter-UP de la variance est sensiblement supérieure à zéro. On trouvera des remarques à ce sujet dans Herzog et Scheuren (1976, p. 398) et Wolter (1985, p. 47). Enfin, Jang et Eltinge (1996) donnent un exemple où la variance intra-UP présente de l'intérêt en soi. Dans certaines applications, on estime la variance intra-UP et les valeurs apparentes selon l'hypothèse apparente que l'estimation est stable, c'est-à-dire présente une variance relativement faible. Nous montrerons ici qu'il peut s'avérer important de vérifier cette hypothèse au moyen de

données et que les principes ordinaires reposant sur le plan d'échantillonnage peuvent déboucher sur des méthodes de contrôle assez simples. Nous insisterons sur les méthodes applicables aux plans d'échantillonnage qui comptent un nombre moyen ou important de strates et un petit nombre d'UPF par strate.

La partie 2.1 passe en revue les estimateurs pertinents de la variance intra-UPF et de la variance globale de la strate. À la partie 2.2, on identifie deux composantes distinctes de la variance de l'estimateur de la variance intra-UPF. Des estimateurs simples de la variance de deux estimateurs de la variance intra-UPF reposant sur le plan de l'échantillonnage sont présentés à la partie 2.3. La partie 3 propose deux mesures liées aux degrés de liberté.

À la partie 4, on montre comment utiliser les méthodes associées au plan d'échantillonnage pour évaluer la stabilité des quantités qui dépendent à la fois de l'estimateur de la variance intra-UPF et de l'estimateur de la variance globale de la strate. On attire principalement l'attention du lecteur sur un estimateur de la variance intra-UPF et sur un estimateur de la variance globale de la strate divisée par la variance intra-UPF. À la partie 4.2, il est question d'un jeu de méthodes articulées sur les mesures de la stabilité de la partie 2 et de quelques hypothèses modérément restrictives à l'égard du moment. Un deuxième jeu de méthodes reposant sur le groupement des strates est présenté à la partie 4.3.





Il est possible d'obtenir les estimations répétées de ces corrélations.

Dans (A1), la covariance devient

$$\begin{aligned} \text{Cov}(X_{+..}, X_{+..}) &= \text{Cov}(X_{1.1} + X_{1.5}, X_{2.2} + X_{2.6}) \\ &= \text{Cov}(X_{1.1}, X_{2.2}) + \text{Cov}(X_{1.1}, X_{2.6}) + \\ &\quad \text{Cov}(X_{1.5}, X_{2.2}) + \text{Cov}(X_{1.5}, X_{2.6}) \\ &= 2(r_1 + \gamma) \text{Var}(X_{1j}), \end{aligned} \quad (\text{A2})$$

par simplification, en supposant que  $\text{Var}(X_{1j})$  est constante pour toutes les valeurs  $j$  et  $j$ . La variance de l'estimation d'un mois complet,  $\text{Var}(\sum_{j=1}^8 X_{1j})$ , est disponible sous forme d'une fonction de variance généralisée (FVG). Nous nous servons de cette estimation pour calculer  $\text{Var}(X_{1j})$  grâce à la dérivation qui suit:

$$\begin{aligned} \text{Var}\left(\sum_{j=1}^8 X_{1j}\right) &= \sum_{j=1}^8 \sum_{k=1}^8 \text{Cov}(X_{1j}, X_{1k}) \\ &= \sum_{j=1}^8 \text{Var}(X_{1j}) + \sum_{j \neq k}^8 \text{Cov}(X_{1j}, X_{1k}) \\ &= (8 + 56\omega) \text{Var}(X_{1j}) \end{aligned}$$

aussi

$$\text{Var}(X_{1j}) = (8 + 56\omega)^{-1} \text{Var}\left(\sum_{j=1}^8 X_{1j}\right). \quad (\text{A3})$$

## BIBLIOGRAPHIE

- ADAMS, D.E. (1991). Current population survey month-in-sample (MIS) bias index research. *Mémorandum interne, Demographic Statistical Methods Division, U.S. Bureau of the Census, Washington, DC.*
- DONNER, A., et LI, K.Y.R. (1990). The relationship between chi-square statistics from matched and unmatched analyses. *Journal of Clinical Epidemiology*, 43, 827-831.
- FAY, R. (1990). VPLX: Variance estimates for complex samples. *Proceedings of the Section on Survey Research Methods, American Statistical Association.*
- FAY, R. (1985). A jackknifed chi-squared test for complex samples. *Journal of the American Statistical Association*, 80, 148-157.
- FEUER, E.J., et KESSLER, L.G. (1989). Test statistic and sample size for a two-sample McNemar test. *Biometrics*, 45, 629-636.
- FISHER, R., ROBINSON, E., THOMPSON, J., et WELCH, M. (1993). Variance estimation in the current population survey overlap test. *Proceedings of the Section on Survey Research Methods, American Statistical Association.*
- FISHER, R., et MCGINNNESS, R. (1993). Correlations and adjustment factors for current population survey. *Mémorandum interne, Demographic Statistical Methods Division, U.S. Bureau of the Census, Washington, DC.*
- HOGUE, C. (1984). History of the problems encountered estimating gross flows. *Proceedings of the Conference on Gross Flows in Labor Force Statistics*, 1-7.
- JABINE, T.B., et SCHEUREN, F.J. (1986). Record linkages for statistical purposes: Methodological Issues. *Journal of Official Statistics*, 2, 3, 255-277.
- McNEMAR, Q. (1947). Note on the sampling error of the differences between correlated proportions of percentages. *Psychometrika*, 12, 153-157.
- MARASCULLO, L.A., OMEICH, C.L., et GOKHALE, D.V. (1988). Planned and post hoc methods for multiple-sample McNemar (1947) tests with missing data. *Psychological Bulletin*, 103, 238-245.
- RAO, J.N.K., et WU, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231-241.
- RAO, J.N.K., et SCOTT, A.J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *The Annals of Statistics*, 12, 46-60.
- SHOEMAKER, H.H. (1993). Results from the current population survey CATI phase-in project. *Proceedings of the Section on Survey Research Methods, American Statistical Association.*
- STASNY, E.A., et FIENBERG, S.E. (1984). Some stochastic models for estimating gross flows in the presence of nonrandom nonresponse. *Proceedings of the Conference on Gross Flows in Labor Force Statistics*, 25-39.
- THOMPSON, J. (1994). Mode effects analysis of labor force estimates. Current Population Survey Overlap Analysis Team Technical Report 3. U.S. Bureau of the Census, Washington, DC.
- THOMPSON, J., et FISHER, R. (1994). Two sample McNemar test for complex surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

#### 4. CONCLUSION

des recherches entreprises par le personnel du Bureau of the Census. Les opinions exprimées dans ce document n'engagent que les auteurs et ne reflètent pas nécessairement le point de vue du Bureau of the Census.

#### ANNEXE

Pour la modification du test de McNemar faisant appel aux données non liées, on estime  $(p_2 - p_1)$  au moyen de  $X_{+,N_1} + X_{+,N_2}$ , où  $X_{+,N_1}$  et  $N_2$  sont des estimations pondérées, et de

$$\widehat{\text{Var}}(p_2 - p_1) = \left[ \frac{N_1}{X_{+,N_1}} \right]_2 \left[ \frac{N_1}{\text{Var}(X_{+,N_1})} - \frac{X_{+,N_1}^2}{N_1^2} \right] + \left[ \frac{N_2}{X_{+,N_2}} \right]_2 \left[ \frac{N_2}{\text{Var}(X_{+,N_2})} - \frac{X_{+,N_2}^2}{N_2^2} \right]$$

$$- \frac{2X_{+,N_1}X_{+,N_2}}{\text{Cov}(X_{+,N_1}, X_{+,N_2})} + \frac{N_1N_2}{\text{Var}(N_2)} \left[ \frac{N_1N_2}{\text{Cov}(X_{+,N_1}, X_{+,N_2})} - \frac{X_{+,N_1}^2}{N_1^2} - \frac{X_{+,N_2}^2}{N_2^2} \right]$$

Dans la présente annexe, nous examinons comment on a dérivé le terme de covariance dans l'estimation de la variance, en ne tenant compte que des données non liées.

Soit la corrélation interne au panel

$$(A1) \quad \text{Cov}(X_{+,N_1}, X_{+,N_2}) = \text{Cov} \left( \sum_{j=1,5}^{j=1,5} X_{1,j}, \sum_{j=2,6}^{j=2,6} X_{2,j} \right)$$

où  $X_{1,j}$  représente le niveau de l'échantillon pondéré pour le mois  $i$  et le mois de l'échantillon (ME)  $j$ . Remarquons que  $X_{1,j}$  et  $X_{2,j+1}$  viennent du même groupe de renouvellement, à moins que  $j = 4$ , puisque le groupe de renouvellement est retiré de l'échantillon pour une période de huit mois après en avoir fait partie pendant quatre mois.

On suppose que les corrélations entre  $X_{1,j}$  et  $X_{2,m}$  se décomposent en trois groupes distincts:

1) Corrélation interne au groupe de renouvellement,

$$\text{Cov}(X_{1,j}, X_{2,i+j+1}) = r_1, \text{ quand } j = 1, 2, 3, 5, 6, 7,$$

2) Corrélation interne au mois entre les groupes de renouvellement,

$$\text{Cov}(X_{1,j}, X_{2,k}) = \omega, \quad k \neq j, \text{ et}$$

3) Corrélation entre les groupes de renouvellement et entre les mois

$$\text{Cov}(X_{1,j}, X_{2,i+1,k}) = \gamma, \quad k \neq j+1 \text{ où } j = 3.$$

Nous avons illustré deux modifications aux tests de McNemar à simple et à double échantillon au moyen des données d'une enquête complexe, et présentée des applications pour la modification faisant intervenir des données non liées. Lorsque l'enquête n'est pas longitudinale, l'application qui fait appel aux données liées créera une structure variance-covariance inconnue et intégrera un facteur de variance attribuable à l'erreur d'appariement. Dans un tel cas, recourir aux données non liées serait sensé, compte tenu de l'interprétation du modèle, même si les résultats obtenus au moyen des estimations (non liées) des probabilités marginales laissent plus à désirer que ceux provenant d'un modèle apparié bien construit. Avec une enquête longitudinale, cependant, la première méthode pourrait s'avérer préférable, car elle constitue le prolongement direct du test classique, donc son interprétation ressemble à celle qu'on retrouve dans les traités.

Le test de McNemar à double échantillon n'est pas la seule approche envisageable dans la situation décrite à la partie 2.2.2. Une autre façon de résoudre la forme non liée du problème consisterait à utiliser un modèle log-linéaire pour un tableau de contingence  $2 \times 2 \times 2$ , comme le décritiven Rao et Scott (1984). Dans l'un ou l'autre cas, des compromis s'imposent. L'interprétation du test de McNemar est intuitive: il s'agit d'un modèle causal ou d'un plan de type expérimental qui exige des mesures répétées. L'interprétation du modèle à tableau de contingence  $2 \times 2 \times 2$  est sans doute moins. Néanmoins, on remarquera que la variable à tester dans le test de McNemar ressemble à celles de Wald, qu'on juge souvent moins efficaces que les statistiques de type chi-carré, lire par exemple Fay (1985). Enfin, il convient de souligner qu'à l'inverse de la formule de Rao-Scott, l'approche décrite ici inclut des dispositions explicites concernant l'usage des données liées.

Parmi les aspects qui pourraient justifier des recherches plus poussées, mentionnons une étude de l'efficacité des épreuves avec les données d'un échantillon complexe, l'estimation de la variance et de la covariance des données couplées, y compris l'apport de la variance des erreurs d'appariement, et la variation de l'efficacité des deux approches. Pour ce qui est des applications à l'analyse des données, le test de McNemar à double échantillon semble présenter une certaine utilité lorsqu'on désire comparer les aspects de différentes méthodes d'enquête ou les effets d'une méthode quelconque sur les réponses dans le temps. Par sa conception, cette approche n'est pas paramétrique; quand l'approximation est valable, elle permet d'éviter les écueils habituellement associés aux tests qui reposent sur l'application d'un modèle.

Les auteurs tiennent à remercier James Hartman, Alfredo Navarro, James Roebuck, Lynn Weidman, l'arbitre et les réviseurs pour leurs commentaires judicieux. Ils remercient aussi Sue Chandler pour avoir accepté de dactylographier ce document. Le présent rapport expose les résultats généraux

#### REMERCIEMENTS



de chômage moyen du panel d'essai parait plus important, ainsi que l'indiquent les résultats significatifs obtenus lors du test de l'hypothèse  $\bar{1} \bar{H} = 0$ .

Les tests  $t$  à double échantillon présentés dans Thompson (1994) avaient aussi permis de déceler une différence positive entre les taux de chômage moyens des panels, à partir des données sur le panel fractionné de la Current Population Survey. Bref, ils montraient que l'ITAOOC relevait le taux de chômage. Ces résultats sont cohérents avec ceux du projet d'intégration de l'ITAOOC à la Current Population Survey rapportés par Shoemaker (1993). L'analyse des données du panel fractionné de la Current Population Survey le confirme. Une fois de plus, on ne peut attribuer entièrement la migration positive nette des non chômeurs vers les chômeurs à l'ITAOOC, les facteurs confondants mentionnés en 3.4.1 affectant aussi le panel d'essai (ITAOOC).

3.5 Discussion

Nos résultats semblent déboucher sur des conclusions contradictoires quant aux effets de l'ITAOOC sur le flux de chômeurs. L'ITAOOC n'a cependant pas le même effet dans les deux tests.

La principale différence réside peut-être dans le questionnaire. Les données de l'enquête parallèle ont été recueillies au moyen du questionnaire de la Current Population Survey, lequel venait d'être remanié. Le nouveau questionnaire a été conçu en vue d'une application automatique. L'ancien questionnaire utilisé dans le cadre du projet d'intégration de l'ITAOOC, par contre, devait être administré comme un questionnaire papier et crayon. Les intervieweurs ont été contraints de mémoriser des enchaînements complexes. Pour alléger au maximum le fardeau du répondant, les deux versions du questionnaire avaient été élaborées en fonction d'une interview de 20 minutes, en moyenne. Le questionnaire automatisé permet à l'intervieweur de recueillir plus de renseignements (et de précisions) dans le même laps de temps, car il n'a plus besoin d'établir le cheminement de l'interview. Outre cette différence attribuable à l'automatisation, les questions sur l'activité ne sont pas libellées de la même façon.

Lors de l'enquête parallèle, on s'est servi du même questionnaire pour les interviews sur le terrain (ordinateur portatif) et celles des installations d'ITAOOC. Le projet d'intégration de l'ITAOOC à la Current Population Survey, par contre, recourait à deux versions de l'ancien questionnaire: une version papier pour les interviews sur place et une version automatisée, avec une question d'introduction légèrement différente sur l'activité, pour l'ITAOOC.

Étant donné ces différences et les avertissements relatifs aux données sur le panel fractionné de l'enquête parallèle, nous qualifions nos résultats de préliminaires. Nous recommandons d'approfondir l'analyse à partir des techniques de McNemar à simple et à double échantillon au moyen des nouvelles données du panel fractionné de la Current Population Survey, qui reposent sur l'ancien plan d'échantillonnage du projet d'intégration de l'ITAOOC et le nouveau questionnaire, entièrement automatisé.

Avant fondé nos prévisions sur les effets du biais attribuable au mois de l'échantillon mentionné dans Adams (1991), nous pensions que l'estimation de  $p_1$  pour le panel témoin (premier et cinquième mois de l'échantillon) serait plus importante que l'estimation de  $p_2$  (correspondante pour le deuxième et le sixième mois. La chose est exacte en général: bien que très variable, l'estimation de  $p_1$  dépasse en moyenne l'estimation de  $p_2$  d'environ 4%. Les deux panels constituant un échantillon représentatif du même échantillon initial, on présume que le mois de l'échantillon introduit le même biais dans les deux panels. La valeur  $p_2$  estimée pour le panel d'essai (ITAOOC) dépasse en moyenne la valeur estimée de  $p_1$ . Compte tenu du comportement des estimations du panel témoin, ces résultats confirment dans une certaine mesure l'existence d'un effet de l'ITAOOC.

Il vaut la peine de souligner le passage de l'état de non-chômeur à celui de chômeur dans le panel d'essai. Cette observation trouve confirmation dans le résultat significatif obtenu lorsqu'on teste l'hypothèse omnibus (valeur prédictive = 0.00) et le résultat significatif obtenu pour l'hypothèse  $\bar{1} \bar{H} = 0$  (valeur prédictive = 0.00). À l'inverse des résultats de l'enquête parallèle, qui apparaissent en 3.4.1, ces données proviennent dans une certaine mesure qu'on obtient un taux de chômage plus élevé avec l'ITAOOC. Les résultats du test de McNemar à double échantillon qui apparaissent au tableau 4 étayent un peu plus cette conclusion. Les résultats mensuels du tableau 4 donnent une indication de la variation du flux de chômage d'un panel à l'autre. Par ailleurs, la valeur du test omnibus est significative (valeur prédictive = 0.00). Le flux

Variable Z	Erreur-type [( $p_2 - p_1$ ) - ( $p_2^* - p_1^*$ )]	Variable Z	Valeur prédictive
10/92 - 11/92	1.18	0.50	0.02
11/92 - 12/92	0.22	0.50	0.67
12/92 - 01/93	-0.29	0.45	0.52
01/93 - 02/93	0.92	0.45	0.04
02/93 - 03/93	-0.10	0.42	0.81
03/93 - 04/93	0.81	0.45	0.07
04/93 - 05/93	0.41	1.01	0.31
05/93 - 06/93	1.16	2.41	0.02
06/93 - 07/93	0.45	1.07	0.28
07/93 - 08/93	0.95	0.46	0.04
08/93 - 09/93	0.14	0.39	0.71
09/93 - 10/93	1.69	0.44	0.00
10/93 - 11/93	0.26	0.40	0.51
11/93 - 12/93	0.40	0.66	

Tableau 4  
Test de McNemar à double échantillon - données non liées de la Current Population Survey

cadre de cette étude s'étaient servis d'une version automatisée de l'ancien questionnaire de la Current Population Survey, dans laquelle la question d'introduction sur la population active avait été légèrement modifiée. On trouvera plus de précisions à ce sujet dans Thompson (1994). Les données examinées ici sont contemporaines à celles du panel fractionné de l'enquête parallèle examinées en 3.4.1. Elles couvrent donc la période d'octobre 1992 à décembre 1993, à l'exclusion de février et de mars 1993. Les cellules du panel d'essai (TTAOC) et du panel témoin (non-TTAOC) devaient comprendre plus d'une centaine d'éléments. Tous les autres mois de données contigus sont donc inclus.

Les résultats du test de McNemar à un échantillon pour les deux panels apparaissent au tableau 3. Les statistiques  $p_1$  et  $p_2$  indiquées correspondent à un pourcentage du nombre estimé de chômeurs par rapport à la population estimée totale du panel.

Comme c'est le cas pour les données sur le panel de l'enquête parallèle, l'application du test de McNemar à un échantillon aux données du projet d'intégration de l'ITAOOC teste l'espérance mathématique que la proportion de chômeurs ne

Tableau 3  
Test de McNemar à un échantillon pour les panels de la Current Population Survey – données non liées

Panel d'essai				Panel témoin					
Période	$p_2 - p_1$	Erreur-type ( $p_2 - p_1$ )	Variable Z	Valeur prédictive	Période	$p_2 - p_1$	Erreur-type ( $p_2 - p_1$ )	Variable Z	Valeur prédictive
10/92 – 11/92	1.13	0.16	7.63	0.00	10/92 – 11/92	0.05	0.47	0.11	0.92
11/92 – 12/92	0.07	0.17	0.44	0.66	11/92 – 12/92	-0.14	0.47	-0.30	0.76
12/92 – 01/93	0.43	0.13	3.46	0.00	12/92 – 01/93	0.72	0.43	1.68	0.09
01/93 – 02/93	0.00	0.14	0.03	0.97	01/93 – 02/93	-0.91	0.43	-2.11	0.03
02/93 – 03/93	-0.25	0.14	-1.81	0.07	02/93 – 03/93	-0.16	0.39	-0.40	0.69
03/93 – 04/93	-0.25	0.14	-1.81	0.07	03/93 – 04/93	-0.16	0.39	-0.40	0.69
04/93 – 05/93	0.63	0.13	4.99	0.00	04/93 – 05/93	-0.18	0.43	-0.42	0.67
05/93 – 06/93	0.88	0.13	6.56	0.00	05/93 – 06/93	0.47	0.38	1.22	0.22
06/93 – 07/93	0.84	0.13	6.49	0.00	06/93 – 07/93	-0.32	0.46	-0.68	0.49
07/93 – 08/93	-0.07	0.14	-0.51	0.61	07/93 – 08/93	-0.52	0.39	-1.32	0.19
08/93 – 09/93	0.42	0.13	3.17	0.00	08/93 – 09/93	-0.54	0.44	-1.21	0.23
09/93 – 10/93	0.06	0.12	0.52	0.60	09/93 – 10/93	-0.08	0.37	-0.22	0.83
10/93 – 11/93	1.05	0.12	8.45	0.00	10/93 – 11/93	-0.63	0.42	-1.50	0.13
11/93 – 12/93	0.18	0.14	1.27	0.20	11/93 – 12/93	-0.09	0.37	-0.23	0.82

Pris séparément, les résultats mensuels ne révèlent pas l'existence d'une variation du flux du chômage entre les deux panels. Par ailleurs, la variable omnibus a une valeur significative (valeur prédictive = 0.00). Le flux de chômage moyen semble inférieur pour le panel d'essai, comme l'indiquent les résultats significatifs du testage de l'hypothèse  $\bar{1} \bar{1} = 0$ , où  $\bar{1}$  est la valeur vectorielle de  $((p_2 - p_1) - (p_2^* - p_1^*))$ , chaque élément correspondant à une estimation mensuelle (valeur prédictive = 0.01).

Dans le cadre de ces tests, nous formulons des remarques au sujet des contrastes dans le tableau des probabilités, en vue de trouver des indicateurs susceptibles de nous éclairer au sujet de l'effet qu'un traitement peut avoir sur la variation du chômage. Tel qu'indiqué précédemment, le mois de l'échantillon introduit un biais dans les tests à un échantillon. Les hypothèses testées portent sur des combinaisons du flux net dans un panel et du biais pour le mois de l'échantillon. La difficulté est quelque peu atténuée dans les tests à double échantillon. En effet, si le biais du mois de l'échantillon s'additionne et affecte de façon identique les deux panels, ses effets s'annuleront. Par ailleurs, l'effet sera légèrement atténué dans le test à double échantillon, même s'il varie entre les deux échantillons ou s'il est multiplicatif. Notre analyse de sensibilité préliminaire révèle que les tests à un échantillon sont sensibles au biais du mois de l'échantillon, mais qu'il n'en va pas autant des tests à double échantillon.

Les tests à double échantillon que présente Thompson (1994) ne permettent pas de déceler une différence entre les taux de chômage moyens des panels, à partir des données du panel fractionné de l'enquête parallèle. Ce résultat contraste avec ceux du projet d'intégration de l'ITAOOC à la Current Population Survey: en deux ans, le panel soumis à l'ITAOOC (trousse expérimentale) a toujours donné un taux de chômage significativement plus élevé que le panel qui ne subissait pas la nouvelle méthode d'interview (groupe témoin). On lira pour cela Shoemaker (1993). Notre propre analyse des données sur le panel fractionné de l'enquête parallèle fournit la preuve que la valeur prévue de la proportion de chômeurs est plus faible avec l'ITAOOC. Les données soulèvent toutefois quelques difficultés. Tout d'abord, tel qu'indiqué précédemment, le panel d'essai (ITAOOC) intègre des facteurs confondants puisque les répondants ne subissent pas tous la deuxième interview d'un centre téléphonique. En deuxième lieu, la taille prévue de l'échantillon dans les cellules du panel témoin approchait dix fois les mois, valeur assez faible pour que la distribution s'effectue de façon imprévisible. Cette dernière difficulté ne pose pas de problème avec le projet d'intégration de l'ITAOOC à la Current Population Survey dont il est

### 3.4.2 Résultats du projet d'intégration de l'ITAOOC à la Current Population Survey

Le projet d'intégration de l'ITAOOC à la Current Population Survey constitue la suite logique à l'étude décrite par Shoemaker (1993). Son objectif principal consistait à jauger l'effet que l'inclusion de l'ITAOOC aura sur le taux de chômage. Les intervieweurs recourant à l'ITAOOC dans le



3.4 Résultats  
3.4.1 Étude du panel fractionné de l'enquête parallèle

Nous n'avons pas déterminé l'efficacité de la moyenne semble se rapprocher le plus de la taille prévue de la cellule. Puisqu'il équivaut à la moyenne géométrique de  $n_1$  et  $n_2$ , qui est impossible d'observer  $n^{(12)}$ , on pré-dénominateur du total prévu de la cellule varie avec la marge pour les deux marges correspondant aux deux mois. Bref, le données marginales non liées. La taille de l'échantillon diffère de contingence est que les cellules devraient au moins comporter cinq éléments. La Current Population Survey et l'enquête parallèle recourant à des plans d'échantillonnage avec grappes, nous avons jugé bon de corriger la valeur limite à la hausse. C'est pourquoi nous l'avons multipliée par l'effet du plan d'échantillonnage. Nous avons aussi majoré la valeur limite de la taille prévue des cellules afin de compenser la corrélation entre le nombre de rangées et de colonnes des tableaux, de façon à obtenir une taille limite finale égale à dix.

Cette partie présente les résultats officiels des tests de McNemar à simple et à double échantillon appliqués aux données du panel fractionné de l'enquête parallèle. Bien que les données aient été recueillies mensuellement, la petite taille prévue des cellules du panel témoin nous a incités à omettre plusieurs fois des mois adjacents. Le tableau 1 présente les statistiques sommaires des tests «mensuels» à un échantillon relatifs à chaque panel, d'après les données non liées du panel et de l'enquête parallèle. Le tableau 2 reproduit les statistiques sommaires des tests à double échantillon reposant sur les données non liées.

Tableau 1  
Test de McNemar à un échantillon pour les panels de l'enquête parallèle – données non liées

Panel d'essai				
Période	$p_2 - p_1$	Erreur-type ( $p_2 - p_1$ )	Variable Z	Valeur prédictive
10/92 – 11/92	-0.62	0.29	-2.18	0.03
11/92 – 12/92	-0.47	0.28	-1.68	0.09
04/93 – 05/93	-0.76	0.27	-2.84	0.00
06/93 – 07/93	-0.04	0.27	-0.16	0.88
08/93 – 09/93	-0.66	0.27	-2.42	0.02

Panel témoin				
$p_2 - p_1$	Erreur-type ( $p_2 - p_1$ )	Variable Z	Valeur prédictive	
10/92 – 11/92	2.44	0.81	3.02	0.00
11/92 – 12/92	0.11	0.83	0.14	0.89
04/93 – 05/93	0.20	0.72	0.27	0.78
06/93 – 07/93	0.97	0.71	1.38	0.17
08/93 – 09/93	-1.73	0.68	-2.54	0.01

Les valeurs déclarées de  $p_1$ ,  $p_2$ ,  $p_1'$ , et  $p_2'$  correspondent à la proportion estimée de chômeurs par rapport à la population estimée totale du panel. On se rappellera que  $p_1$  et  $p_1'$

On s'attendait à ce que les estimations présentent un certain biais pour le mois de l'échantillon. En effet, Adams (Bureau of the Census 1991) a constaté que l'estimation de  $p_1$  d'après le premier et le cinquième mois de l'échantillon de la Current Population Survey globale dépassait les valeurs analogues du deuxième et du sixième mois de l'échantillon ( $p_2 - p_1$ ), qui repose sur les données de la Current Population Survey, est donc négative. Comme on peut le constater au tableau 1, tel n'est pas le cas avec les estimations du panel témoin de l'enquête parallèle: contrairement à ce qu'on pourrait croire, l'écart estimé ( $p_2' - p_1'$ ) est habituellement positif. Peut-être le doit-on à une variation dans le temps, la région ou le plan d'échantillonnage. Adams s'est servi des données de la Current Population Survey de 1987 pour estimer le biais associé aux groupes de renouvellement à l'échelon national. Dans chaque test à un échantillon, au flux net se mêle donc un effet non quantifié, attribuable au biais du mois dans l'échantillon.

On remarquera le flux négatif du chômage pour le panel d'essai. La valeur significative obtenue avec le test officiel d'essai. La valeur significative obtenue avec le test officiel pour l'hypothèse omnibus (valeur prédictive = 0.00) et la valeur significative de l'hypothèse  $\bar{H} = 0$  (valeur prédictive = 0.00) le confirment.

Voici les résultats obtenus avec le test de McNemar à double échantillon.

Tableau 2  
Test de McNemar à double échantillon – données non liées de l'enquête parallèle

Période	( $p_2 - p_1$ ) - Erreur-type [( $p_2 - p_1$ ) - ( $p_2' - p_1'$ )]	Variable Z	Valeur prédictive
10/92 – 11/92	-3.06	0.86	-3.58
11/92 – 12/92	-0.58	0.88	-0.66
04/93 – 05/93	-0.95	0.77	-1.24
06/93 – 07/93	-1.02	0.76	-1.34
08/93 – 09/93	1.08	0.74	1.47
			0.14
			0.18
			0.22
			0.51
			0.00

servi de son logiciel VPLX (Fay 1990) pour effectuer une deuxième estimation de la corrélation entre les groupes de renouvellement des chômeurs et de la population active civile, à partir des données de la Current Population Survey de septembre 1992 à décembre 1993. Nous avons utilisé cette corrélation pour les variables à tester reposant sur les données non liées, en supposant qu'elles ne présenteraient pas de différence entre les enquêtes (Current Population Survey et enquête parallèle) ou les régions (données nationales et infranationales). De là, nous avons dérivé la corrélation intérieure au panel pour la population civile en raccordant les autocorrélations calculées précédemment (Fisher et McGinness 1993) et les estimations de la variance aux estimations de chaque groupe de renouvellement. On trouvera plus de détails sur l'estimation des corrélations dans l'annexe.

Nous n'avons pas recouru à la modification avec couplage pour diverses raisons, la principale étant qu'il est difficile de coupler de façon longitudinale les données. Par ailleurs, nous n'avons pu évaluer l'efficacité du couplage. Enfin, nous ne possédions pas d'estimation de la corrélation pour les données liées.

L'hypothèse que la probabilité d'une non-réponse (ou d'un non-appariement) est aléatoire est implicite à l'analyse des données non liées. Nous supposons que la probabilité d'une non-réponse lors d'un mois quelconque ne dépend pas de la classification du répondant au sein de la population active le mois précédent. Cette hypothèse n'est pas acceptée partout. De fait, Slasny et Fienberg (1984) soutiennent le contraire et proposent plusieurs autres modèles à temps fini afin de permettre l'utilisation des données de la CPS liées longitudinalement. Dans notre application, l'estimation des probabilités marginales reposant sur les données liées (éventuellement) mal appariées était presque identique à l'estimation dérivant des données non liées. C'est pourquoi nous pensons que l'analyse n'a pas beaucoup souffert de notre hypothèse.

### 3.3 Diagnostic

La petite taille de l'échantillon qu'on prévoit dans les cellules suscitera des résultats très variables, sur lesquels on ne pourra se fier. Nous ne connaissons pas de méthode générale permettant de calculer des échantillons de taille adéquate pour ce genre d'analyse à partir des données d'une enquête complexe. Nous avons naïvement recouru à une version légèrement modifiée du test chi-carré classique de diagnostic de Pearson pour obtenir une valeur limite de la façon suivante.

Ainsi qu'on l'indique à la partie 2.2.2, soit

$x_{+i}$  = le nombre de chômeurs non pondéré au mois  $i$ ;  
 $x_{-i}$  = le nombre de personnes occupées non pondéré au mois  $i$ ;

$x_{+j}$  = le nombre de chômeurs non pondéré au mois  $j$ ;  
 $x_{-j}$  = le nombre de personnes occupées non pondéré au mois  $j$ ;

au mois 2.

Rappelons que dans le tableau de contingence habituel,  $E[+ -] = x_{+i} \cdot x_{-j} / n^{(12)}$ ,  $E[- +] = x_{-i} \cdot x_{+j} / n^{(12)}$  selon l'hypothèse d'indépendance (sans tenir compte des valeurs manquantes). Pour estimer la taille prévue des cellules, nous utilisons les

consécutifs, l'interview est interrompue pendant les huit mois suivants, puis on la reprend quatre mois d'affilée. La première et la cinquième interviews se font sur place, mais les autres sont effectuées au téléphone, dans la mesure du possible. La première et la cinquième interviews produisent donc une valeur de base de la population active; par ailleurs, la deuxième et la sixième permettent de mesurer l'activité

«après traitement».

Pour créer le panel nécessaire aux deux études, on a divisé de façon aléatoire l'échantillon recueilli dans les régions sélectionnées en deux panels représentatifs, par une méthode d'échantillonnage systématique. Les sujets du panel d'essai ont été jugés admissibles à l'ITAO, bref les ménages de l'échantillon ont subi l'interview à partir des installations centrales, après l'interview initiale (première et cinquième). Pour subir l'ITAO, le répondant devait avoir le téléphone et parler l'anglais ou l'espagnol, et accepter de passer l'interview téléphonique au cours des mois subséquents. Les ménages du panel n'ont pas tous reçu une ITAO. Les ménages de l'autre panel faisaient office de témoin.

Le taux de chômage moyen est la principale statistique publiée consécutivement à l'obtention des données de la Current Population Survey. Ce taux correspond au nombre estimé de chômeurs divisé par le nombre estimé de personnes constituant la population active civile (le dénominateur exclut le personnel militaire, les personnes de moins de 16 ans et celles qui ne cherchent plus de travail ou qui ont pris leur retraite). Notre principal objectif était de comprendre comment l'inclusion de l'ITAO pouvait agir sur la probabilité d'un changement au sein de la population active soit, dans le cas qui nous intéresse, d'un passage du chômage au travail (ou vice-versa). Nos statistiques des tests de McNemar à simple et à double échantillon reposaient sur le ratio entre le chômage et la population, plutôt que sur le taux de chômage. De cette façon, on obtient une estimation légèrement plus précise de la proportion en atténuant la variabilité de la variable à tester.

### 3.2 Estimations

L'estimation de chaque mois ou panel est une estimation non biaisée. En d'autres termes, les poids utilisés pour obtenir les estimations ne dépendent strictement que de la probabilité de sélection: chaque poids est le produit du poids de base (inverse de la probabilité d'être sélectionnée pour une UPE), du coefficient de contrôle de la pondération (ajustement pour le sous-échantillonnage sur le terrain) et d'un facteur pour le fractionnement du panel (ajustement pour la probabilité de faire partie d'un panel fractionné). Le coefficient du panel fractionné de l'enquête parallèle est constant par définition: neuf dixièmes de l'échantillon ont été affectés au hasard au panel d'essai. Les coefficients du panel concernant le projet d'intégration de l'ITAO à la CPS, par contre, ne l'étaient pas: l'échantillon du panel d'essai variait mensuellement, une plus grande partie de l'échantillon étant attribuée de façon aléatoire aux installations d'ITAO.

La variance des niveaux a été calculée au moyen de fonctions de variance généralisées (FVG). Pour plus de précisions, on lira Fisher et ses collaborateurs (1993). Robert Fay s'est



On peut donc désormais tester aisément les hypothèses linéaires générales de la forme  $K\bar{\mu}$ . Peut-être voudra-t-on tester l'existence d'un contraste par période, par exemple comparer la différence moyenne de janvier à juin aux données du reste de l'année. L'épreuve sans doute la plus intéressante (pour nos applications) est celle de l'hypothèse  $H_0: \bar{1}\bar{\mu} = 0$ , où  $\bar{1}$  représente l'espérance mathématique d'un des vecteurs déjà décrits.

Un autre test intéressant consiste à tester «l'hypothèse omnibus», c'est-à-dire tester  $H_0: \bar{\mu} = \bar{0}$ . Dans ce cas, les variables à tester sont  $\bar{\lambda}_T^T \sum_{\lambda(1)}^{-1} \bar{\lambda}_T$ ,  $\bar{\lambda}_C^C \sum_{\lambda(1)}^{-1} \bar{\lambda}_C$  et  $\bar{\lambda}_6 \sum_{\lambda(6)}^{-1} \bar{\lambda}_6$ , chacune étant caractérisée par une distribution chi-carré approximative avec  $r$  degrés de liberté,  $r$  correspondant à la dimension du vecteur auquel on s'intéresse.

### 3. APPLICATIONS

Nous appliquerons maintenant les techniques de McNemar à simple et à double échantillon décrites en 2.2.2 et 2.3 pour les données non liées à deux ensembles distincts de données: les données du panel fractionné de l'enquête parallèle à la Current Population Survey et les données du projet d'intégration de l'ITAOOC à la Current Population Survey. Les tableaux 1 et 2 (partie 3.4.1) donnent les résultats pour le panel fractionné de l'enquête parallèle et les tableaux 3 et 4 (partie 3.4.2), ceux relatifs au projet d'intégration de l'ITAOOC à la Current Population Survey.

#### 3.1 Contexte

Les estimations mensuelles officielles de la population active civile obtenues à partir de janvier 1994 reposent sur les données d'une Current Population Survey profondément remaniée. La restructuration de l'enquête prévoyait l'application d'un nouveau questionnaire, entièrement informatisé, et un plus grand recours à l'interview téléphonique assistée par ordinateur centralisée (ITAOOC). Pour mieux jauger les effets de la restructuration sur les chiffres publiés, on a entrepris une enquête parallèle avec le nouveau questionnaire et les nouvelles méthodes de collecte des données, de juillet 1992 à décembre 1993. Outre l'enquête parallèle et la Current Population Survey, des études spéciales ont été intégrées au projet durant la même période et devaient servir à recueillir les données permettant de tester les hypothèses quant aux effets des nouvelles méthodes sur l'estimation de la population active: l'étude du panel fractionné de l'enquête parallèle et le projet d'intégration de l'ITAOOC à la Current Population Survey (un prolongement de l'étude décrite dans Shoemaker 1993).

On s'est particulièrement intéressé aux conséquences d'un recours plus important à l'interview téléphonique assistée par ordinateur centralisée. L'étude décrite dans Shoemaker (1993) avait révélé que la centralisation des interviews téléphonique a tendance à entraîner une surestimation du taux de chômage. Le test de McNemar à double échantillon semblait bien se prêter à l'étude de ce phénomène. Dans le cadre de la Current Population Survey et de l'enquête parallèle, les ménages subissent une interview pendant quatre mois

Si l'enquête est conçue pour recueillir des données longitudinales, pareille modification constitue le prolongement naturel de la méthode que décrivent Feuer et Kessler. On suppose qu'une enquête de ce type intègre un mécanisme efficace permettant de raccorder les sujets d'un mois à l'autre. Souvent toutefois, cela n'est pas le cas et on doit physiquement coupler un ensemble de données à un autre. Par conséquent, les  $n^{(12)}$  éléments du domaine comprendront quelques faux appariements et excluront par accident des appariements véritables. Le poids des enregistrements et l'estimation de la variance devront être corrigés pour tenir compte de l'appariement. Jabin et Scheuren (1986) résument très bien les problèmes méthodologiques que pose l'utilisation de données couplées, tant pour les techniques de couplage des enregistrements faisant appel à un modèle et que pour les techniques spéciales («couplage forcé»).

#### 2.2.2 Deuxième modification: données non couplées

Dans cette méthode, on saute l'étape du couplage longitudinal, car le test de McNemar classique peut être construit au moyen des estimations des probabilités marginales. On présume que, d'après l'hypothèse nulle, la valeur prévue de  $(p_* - p_*)$  est zéro. Marascuilo et ses collaborateurs (1988) décrivent cette situation pour un échantillonnage aléatoire simple.

Les données du premier mois appartiennent au domaine  $M^{(12)} \cup M^{(10)}$ , qui comporte  $n^{(12)} + n^{(10)} = n_1$  éléments, et celles du deuxième mois à  $M^{(12)} \cup M^{(02)}$ , qui renferme  $n^{(12)} + n^{(02)} = n_2$  éléments.

La variable à tester construite à partir des données non couplées est

$$Z_1 = \frac{\sqrt{\text{Var}(p_2 - p_1)}}{p_2 - p_1}$$

$$p_1 = \frac{x_+}{x_-}, \quad p_2 = \frac{u_+}{u_-}$$

Si les deux panels sont indépendants, la variable à tester est

$$Z = \frac{\sqrt{\text{Var}(p_2 - p_1) + \text{Var}(p_2' - p_1')}}{(p_2 - p_1) - (p_2' - p_1')}$$

où

$$p_1' = \frac{x_+'}{x_-'}, \quad p_2' = \frac{u_+'}{u_-'}$$

Comme c'est le cas pour l'application décrite en 2.2.1, toutes les estimations sont pondérées et les variances correspondent à celles de l'enquête complexe.

#### 2.3 Combinaisons linéaires

On peut se servir de la matrice des covariances estimatives pour vérifier les combinaisons linéaires de  $\bar{\lambda}_T$ ,  $\bar{\lambda}_C$  et  $\bar{0}$  dans le temps, où  $\bar{\lambda}_T = \bar{p}_2 - \bar{p}_1$ ,  $\bar{\lambda}_C = \bar{p}_2' - \bar{p}_1'$ , et  $\bar{0} = \bar{\lambda}_T - \bar{\lambda}_C$ , et  $\bar{p}_1, \bar{p}_2, \bar{p}_1'$  et  $\bar{p}_2'$  sont des vecteurs incluant les probabilités marginales pour la période à l'étude.

**Panel d'essai**

Mois 2

Application de la nouvelle technique

Mois 1	+	$x_{1+}$	$x_{1-}$	$x_{1\cdot}$
Pas d'application	-	$x_{2+}$	$x_{2-}$	$x_{2\cdot}$
	*	$x_{+}$	$x_{-}$	$n$

**Panel témoin**

Mois 2

Pas d'application

Mois 1	+	$x'_{1+}$	$x'_{1-}$	$x'_{1\cdot}$
Pas d'application	-	$x'_{2+}$	$x'_{2-}$	$x'_{2\cdot}$
	*	$x'_{+}$	$x'_{-}$	$n'$

Pour chaque panel, soit

$M_{(12)}$ , l'ensemble de cas pour lesquels on obtient une réponse le mois 1 et le mois 2 (cas appariés). Cet ensemble renferme  $n_{(12)} = (x_{++} + x_{+-} + x_{-+} + x_{--})$  éléments;

$M_{(10)}$ , l'ensemble de cas pour lesquels on obtient une réponse le mois 1, mais pas le mois 2. Cet ensemble renferme  $n_{(10)} = (x_{+*} + x_{-*})$  éléments;

$M_{(02)}$ , l'ensemble de cas pour lesquels on obtient une réponse le mois 2, mais pas le mois 1. Cet ensemble renferme  $n_{(02)} = (x_{*+} + x_{*-})$  éléments.

Souignons que  $n$  représente la taille de l'échantillon et  $n'$  est pas pondéré.

Examinons d'abord le cas à un échantillon. Habituellement, le test de McNemar à un échantillon repose sur les réponses appariées  $n_{(12)}$  et  $n'_{(12)}$ , où le caractère prime (') désigne le panel témoin. Dans le scénario à un échantillon, on teste l'hypothèse suivante:

$$H_0: p_{+-} = p_{+*}, \text{ où } p \text{ correspond à la probabilité des cellules}$$

$$H_1: \text{Non } H_0$$

c'est-à-dire l'hypothèse qu'il n'y a pas de passage d'un état à l'autre (+ à -, ou - à +). Un mouvement de ce genre est aussi appelé flux.

Le test à un échantillon peut s'avérer utile comme outil de diagnostic dans le cas à double échantillon. On examine les estimations du panel témoin pour s'assurer qu'il n'y a aucun flux. Un déplacement significatif au sein du panel d'essai peut ensuite être considéré comme un écart par rapport au flux nul ou comme une variation de la probabilité d'obtenir un «+».

L'hypothèse à double échantillon est la suivante:

$$H_0: (p_{+-} - p_{+*}) = (p'_{+-} - p'_{+*})$$

$$H_1: \text{Non } H_0$$

Ces résultats sont valables peu importe le plan d'échantillonnage. Pour appliquer ces équations à une enquête complexe, on recourt aux estimations pondérées ainsi qu'aux variances d'une enquête complexe, au lieu d'utiliser la variance d'un échantillon aléatoire simple.

où

$$Z^* = \frac{\sqrt{\text{Var}(p_{2\cdot}^1 - p_{2\cdot}^2) + \text{Var}(p_{2\cdot}^1 - p_{2\cdot}^2)}}{(p_{2\cdot}^2 - p_{2\cdot}^1) - (p_{2\cdot}^2 - p_{2\cdot}^1)},$$

$$p_{2\cdot}^1 = \frac{n_{(12)}^1}{x_{++}^1 + x_{+-}^1}, \quad p_{2\cdot}^2 = \frac{n_{(12)}^2}{x_{++}^2 + x_{+-}^2},$$

où

$$Z^*_1 = \frac{\sqrt{\text{Var}(p_{2\cdot}^1 - p_{2\cdot}^1)}}{p_{2\cdot}^2 - p_{2\cdot}^1},$$

où  $p_{2\cdot}^1$  représente la probabilité marginale d'obtenir une réponse + au mois 2, compte tenu d'un appariement entre les deux mois, et  $p_{2\cdot}^1$  correspond à la probabilité marginale d'obtenir une réponse + au mois 1, lorsqu'il y a appariement des deux mois.

Pour le test à un échantillon, la variable à tester construite à partir des données du panel est

$$[p_{-+} - p_{+-}] = [(p_{+-} + p_{-+}) - (p_{+-} + p_{-+})]$$

$$= [p_{*-} - p_{*-}]$$

$$= p_{2\cdot}^1 - p_{2\cdot}^1$$

probabilités des cellules et on remarque que

Cette méthode est une application directe du test de McNemar à double échantillon reposant sur les données d'une enquête complexe couplées de façon longitudinale.

**façon longitudinale****2.2.1 Première modification: données couplées de****2.2 Modifications à une enquête complexe**

Le test de McNemar à double échantillon généralisé par Feuer et Kessler (1989) (décrit en 2.2.1 ci-dessous) se borne aux ensembles de valeurs appariées  $M_{(12)}$  et  $M'_{(12)}$ . On peut toutefois ajouter une supposition, afin que les réponses non appariées entrent dans le calcul de la variable à tester. Cette supposition est à l'origine de la discussion de la partie 2.2.2.

En d'autres termes, la différence entre les probabilités que le déplacement se fasse dans un sens ou dans l'autre reste la même, peu importe le traitement. Bref, la différence entre les flux des deux panels est égale à zéro.



# Application des tests de McNemar à l'étude du panel partiel de la Current Population Survey

KATHERINE JENNY THOMPSON et ROBIN FISHER<sup>1</sup>

## RÉSUMÉ

Les résultats des études du panel partiel de la Current Population Survey révèlent que l'interview téléphonique assistée par ordinateur centralisée (TTAOC) a un effet sur l'estimation de la population active. Une hypothèse est que l'ITTAOC accroît la probabilité que le répondant modifie sa déclaration au sujet de sa situation vis-à-vis de l'activité. Le test de McNemar à double échantillon permet de tester cette hypothèse: celle qui nous intéresse est que les changements marginaux relevés dans les tableaux des deux échantillons indépendants sont égaux. Les auteurs présentent deux variantes du test adaptées aux données d'enquêtes complexes, ainsi que leur application aux données du panel partiel de l'enquête parallèle à la Current Population Survey et à celles du projet d'intégration de l'ITTAOC à la Current Population Survey.

MOTS CLÉS: Current Population Survey; enquête parallèle; statistiques non paramétriques.

## 1. INTRODUCTION

Les résultats de l'étude du panel fractionné de l'enquête parallèle à la Current Population Survey et du projet d'intégration de l'ITTAOC à la Current Population Survey semblent indiquer que l'interview téléphonique assistée par ordinateur centralisée (TTAOC) influe sur l'estimation mensuelle de la population active aux États-Unis (Thompson 1994 et Shoemaker 1993). Une hypothèse est que l'ITTAOC accroît la probabilité que le répondant modifie sa déclaration concernant sa situation vis-à-vis de l'activité entre la première interview (sur place) et la seconde (ITTAOC).

On peut recourir au test de McNemar à double échantillon pour vérifier cette hypothèse. Le test de McNemar (1947) a été généralisé à un cas à double échantillon selon l'hypothèse que les changements marginaux observés dans les tableaux 2 x 2 de deux échantillons indépendants sont égaux (Feuer et Kessler 1989). Cette application se rapportait à l'analyse d'une cohorte à double échantillon et supposait un échantillonage aléatoire simple.

Avant qu'on puisse effectuer le test de McNemar, certaines modifications doivent être apportées à la variable à tester quand les données viennent d'une enquête complexe. Tout d'abord, les données n'étant pas recueillies par un échantillonage aléatoire simple et étant pondérées, une estimation distincte de la variance s'impose. Deuxièmement, on doit établir un rapprochement distinct entre les personnes interrogées pour deux mois de données consécutifs, à moins qu'il ne s'agisse d'une enquête longitudinale. En règle générale, un tel rapprochement inclura quelques faux appariements et exclura certains appariements véritables. Cela crée donc une nouvelle source de variance.

Nous proposons deux adaptations à ce test pour les données d'une enquête complexe. Plus précisément, nous

présentons les tests modifiés et leur application à l'étude du panel fractionné de l'enquête parallèle à la Current Population Survey et au projet d'intégration de l'ITTAOC à la Current Population Survey. La partie 2 décrit les modifications apportées au test et fournit des renseignements de base sur les tests de McNemar à simple et à double échantillon (partie 2.1), indique les changements que nécessitent les données des enquêtes complexes (partie 2.2) et comprend certaines remarques sur les applications aux données de plusieurs mois (partie 2.3). La partie 3 expose l'application des tests aux données de l'enquête parallèle à la Current Population Survey et aux données du projet d'intégration de l'ITTAOC à la même enquête. S'y ajoutent des renseignements de base sur les deux études (partie 3.1), des détails sur les estimations du panel et de la variance (partie 3.2), les outils de diagnostic (partie 3.3) et les résultats (partie 3.4). L'article se termine par les conclusions, à la partie 4. Enfin, l'annexe donne des précisions sur l'estimation de la covariance.

## 2. TEST ET MODIFICATIONS

### 2.1 Généralités

Un échantillon est divisé de façon aléatoire pour former deux échantillons représentatifs indépendants (panels tracés). Une fois que les valeurs de base ont été obtenues pour les deux panels, on applique une nouvelle technique à l'un d'eux, l'autre servant de témoin. Les enregistrements sont rapprochés de façon longitudinale après la deuxième mesure. L'appariement peut être positif ou négatif, ou être absent (+, -, ou \*). Puisque les données sont couplées, la cellule «\*\*» reste vide. On peut illustrer visuellement ce scénario de la façon suivante:

<sup>1</sup> Katherine Jenny Thompson, Economic Statistical Methods and Programming Division, et Robin Fisher, Housing and Household Economic Statistics Division, United States Bureau of the Census, Washington, DC 20233, U.S.A.





conditions 4, 6 et 7, on obtient le résultat souhaité en utilisant l'instrument de Cramér-Wold.

**Preuve du Théorème 2.1 (2).** Comme il pourrait y avoir d'autres valeurs  $(\alpha', \beta') \in B_n$  pour lesquelles  $\hat{y}(\alpha', \beta') = \hat{y}(\alpha, \beta)$  pour certains  $(\alpha, \beta) \in B_n$ ,  $G_n$  est toujours une approximation conservatrice.

## BIBLIOGRAPHIE

- BICKEL, P.J., et FREEDMAN, D.A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *The Annals of Statistics*, 12, 470-482.
- BOX, G.E.P., et TIDWELL, P.W. (1962). Transformations of the independent variables. *Technometrics*, 4, 531-550.
- BOX, G.E.P., et COX, D.R. (1964). An analysis of transformations. *Journal of The Royal Statistical Society, Série B*, 26, 211-243, discussions 244-252.
- BREIMAN, L., et FRIEDMAN, J. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80, 580-597.
- CARROLL, R.J., et RUPPERT, D. (1981). On prediction and power transformation family. *Biometrika*, 68, 609-615.
- CARROLL, R.J., et RUPPERT, D. (1988). *Transformation and Weighing in Regression*. London: Chapman and Hall.
- CALVIN, J.A., et SEDRANSKY, J. (1991). Bayesian and frequentist predictive inference for the patterns of care studies. *Journal of the American Statistical Association*, 86, 36-54.
- CHEN, J., et QIN, J. (1992). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika*, 80, 107-116.
- COCHRAN, W.G. (1977). *Sampling Techniques*. (3ième éd.) New York: John Wiley.
- DE VEAUX, R.D., et STEELE, J.M. (1989). ACE guided-transformation method for estimation of the coefficient of soil-water diffusivity. *Technometrics*, 31, 91-98.

- DEVILLE, J., et SÄRNDAAL, C. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- ELLIOTT, J.M. (1977). Statistical analysis of samples of benthic invertebrates. *Freshwater Biological Scientific Publication*, n° 25, (2ième éd.).
- HÄJEK, J. (1960). Limiting Distributions in Simple Random Sampling From a Finite Population. Publications in Mathematics of the Hungarian Academy of Science, 5, 361-374.
- NELDER, J.A., et PREGIBON, D. (1987). An extended quasi-likelihood function. *Biometrika*, 74, 221-232.
- OWEN, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75, 237-249.
- OWEN, A. (1990). Empirical likelihood confidence regions. *The Annals of Statistics*, 18, 90-120.
- PANKRATZ, A., et DUDLEY, U. (1987). Forecasts of power-transformed series. *Journal of Forecasting*, 6, 239-248.
- ROYALL, R.M., et CUMBERLAND, W.G. (1981a). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, 76, 66-88.
- ROYALL, R.M., et CUMBERLAND, W.G. (1981b). The finite-population linear regression estimator and estimators of its variance - An empirical study. *Journal of the American Statistical Association*, 76, 924-930.
- ROYALL, R.M., et CUMBERLAND, W.G. (1985). Conditional coverage properties of finite population confidence intervals. *Journal of the American Statistical Association*, 80, 355-359.
- SÄRNDAAL, C., SWENSSON, B., et WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- SCOTT, A., et WU, C.F. (1981). On the asymptotic distribution of the ratio and regression estimators. *Journal of the American Statistical Association*, 76, 98-102.
- WHITMORE, G.A. (1983). A regression method for censored inverse-gaussian data. *The Canadian Journal of Statistics*, 11, 305-315.

Tableau 2  
Résultats des simulations basées sur 10,000 échantillons aléatoires simples dont la taille est égale à 32

Cancer		Villes		Comtés 60		Comtés 70		Hôpitaux		Ventes	
Ratio	3.26	3.65	3.05	2.90	3.62	2.94	Méthode de régression (variance de régression)	Ncr	0.141	0.116	0.146
Ncr	0.141	0.116	0.146	0.271	0.098	0.176		Méthode de régression (variance jackknife)	Ratio	4.03	3.88
Ncr	0.081	0.102	0.083	0.192	0.068	0.079	Méthode de transformation (toutes les valeurs de x sont connues)	Ratio	5.08	4.00	3.75
Ratio	5.12	3.74	3.37	3.69	4.15	4.90	Méthode de vraisemblance empirique (seul x est connue)	Ncr	0.018	0.074	0.053
Ncr	0.017	0.082	0.081	0.082	0.037	0.006	Méthode de vraisemblance empirique (population créée)	Ratio	3.92	3.92	3.97
Ratio	3.92	3.92	3.97	3.96	3.90	3.99		Ncr	0.057	0.059	0.055

Nous utilisons la transformation logarithmique dans certaines de nos discussions, car il s'agit probablement de la transformation utilisée le plus fréquemment en pratique. Néanmoins, des méthodes plus objectives existent pour choisir le type de transformation. L'une d'elles est la transformée exponentielle bien connue de Box-Cox que nous avons déjà mentionnée; consulter Box et Cox (1964), Box et Tidwell (1962), Carroll et Ruppert (1988). Une autre méthode récente se fonde sur un calcul appelé «espérance conditionnelle alicornative» (Breiman et Friedman 1985, DeVeaux et Steele 1989).

Il existe d'autres moyens d'améliorer le taux conditionnel de couverture. Un de ces moyens consiste à employer des distributions asymétriques d'erreur telles que la famille des distributions gaussiennes inverses (Whitmore 1983). Un autre consiste à adapter les fonctions de quasi-vraisemblance (Nelder et Pregibon 1987) aux problèmes que posent les populations finies.

L'étude de simulation qui suit prouve aussi la validité de notre nouvelle méthode. Pour chacune des six populations réelles, nous créons une nouvelle population en remplaçant les valeurs originales  $y_i$  par

$$y_i^* = \exp\{\alpha + \beta \log(x_i) \theta e_i\},$$

où  $\alpha$ ,  $\beta$  et  $\theta$  sont les estimations des paramètres destinés à ajuster le modèle (2.2), avec  $h = g = \log$ , à l'ancienne population, et où les  $e_i$  sont générées à titre de variables normales réduites indépendantes et distribuées de façon identique. En nous servant des six populations créées qui sont

fixes, nous répétons les simulations de la section 3 unique-ment pour le cas où on connaît  $x$ . Cette étude de simulation est résumée au tableau 2 et le graphique de non-couverture pour la population Comtés 70 figure dans le coin inférieur gauche de la figure 3. (Les graphiques de non-couverture pour d'autres populations sont fort semblables à ce graphique). Cette étude indique clairement que, quand la population finie est générée au moyen d'un modèle de superpopulation tel que (2.2) dont la distribution d'erreur est normale, notre nouvelle méthode produit les taux conditionnels de couverture corrects. Qui plus est, quand nous diminuons la corrélation entre  $x$  et  $y$  jusqu'à une valeur aussi faible que 0,5 pour chacune des six populations en augmentant  $\theta$  et que nous répétons les simulations susmentionnées, les résultats sont aussi bons que ceux présentés au tableau 2 et à la figure 3.

Nous ne considérons que la méthode d'échantillonnage aléatoire simple dans le présent article, mais la méthode proposée est applicable à condition (i) qu'il existe une corrélation linéaire entre  $h(y)$  et  $g(x)$  pour certaines fonctions monotones  $h$  et  $g$ , et (ii) qu'on puisse trouver  $F_N(u)$  ou  $F_N^*(u)$ . Puisque les six populations ont été choisies en s'assurant qu'elles soient représentatives, nous prévoyons que notre nouvelle méthode facilitera l'étude d'autres populations finies.

## REMERCIEMENTS

Nous remercions l'examinateur et le rédacteur dont les commentaires nous ont permis d'améliorer la présentation. Les auteurs sont tous deux titulaires d'une bourse du Conseil de recherche en sciences naturelles et en génie du Canada.

## ANNEXE

**Preuve du Théorème 2.1 (1).** Pour  $n$  importe quels nombres réels  $t_1$  et  $t_2$ , nous avons

$$t_1(\alpha - \alpha_N) + t_2(\beta - \beta_N) =$$

$$t_1 n^{-1} \sum_{i \in S} e_i + \frac{t_2 - t_1 \bar{u}_s}{t_2 - t_1 \bar{u}_s} \sum_{i \in S} (u_i - \bar{u}_s) e_i.$$

En vertu des conditions 1, 2 et 3, nous avons

$$\bar{u}_s - \bar{u}, \quad n^{-1} \sum_{i \in S} (u_i - \bar{u}_s)^2 - \sigma_u^2.$$

Par conséquent, nous pouvons écrire

$$t_1(\alpha - \alpha_N) + t_2(\beta - \beta_N) =$$

$$t_1 n^{-1} \sum_{i \in S} e_i + \frac{\sigma_u^2}{t_2 - t_1 \bar{u}_s} n^{-1} \sum_{i \in S} (u_i - \bar{u}_s) e_i + o_p(n^{-1/2}).$$

La condition de Lindeberg-Hájek est satisfaite pour  $t_1 e_i + t_2 - t_1 \bar{u}/\sigma_u^2 (u_i - \bar{u}) e_i$  quand la condition de moment est 5, consulter Hájek (1960), Scott et Wu (1981); et Bickel et Freedman (1984). En l'appliquant en même temps que les



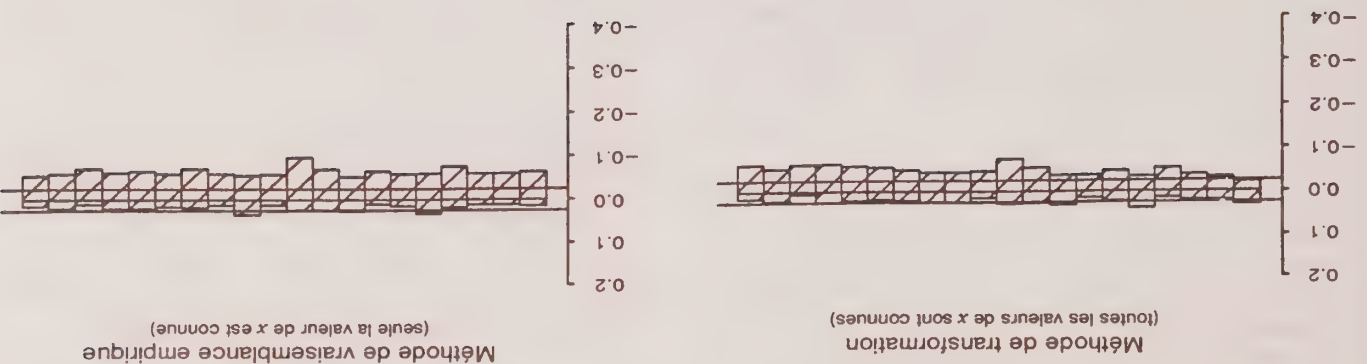
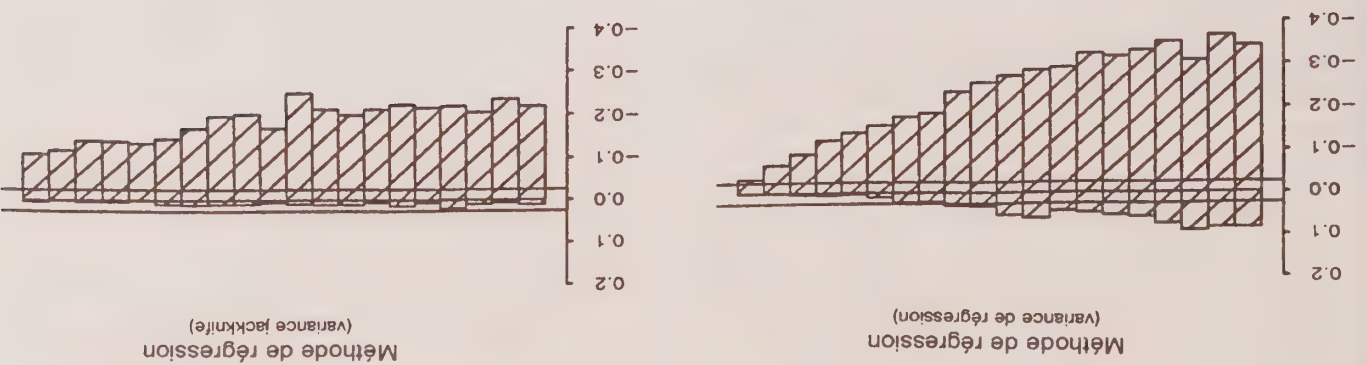
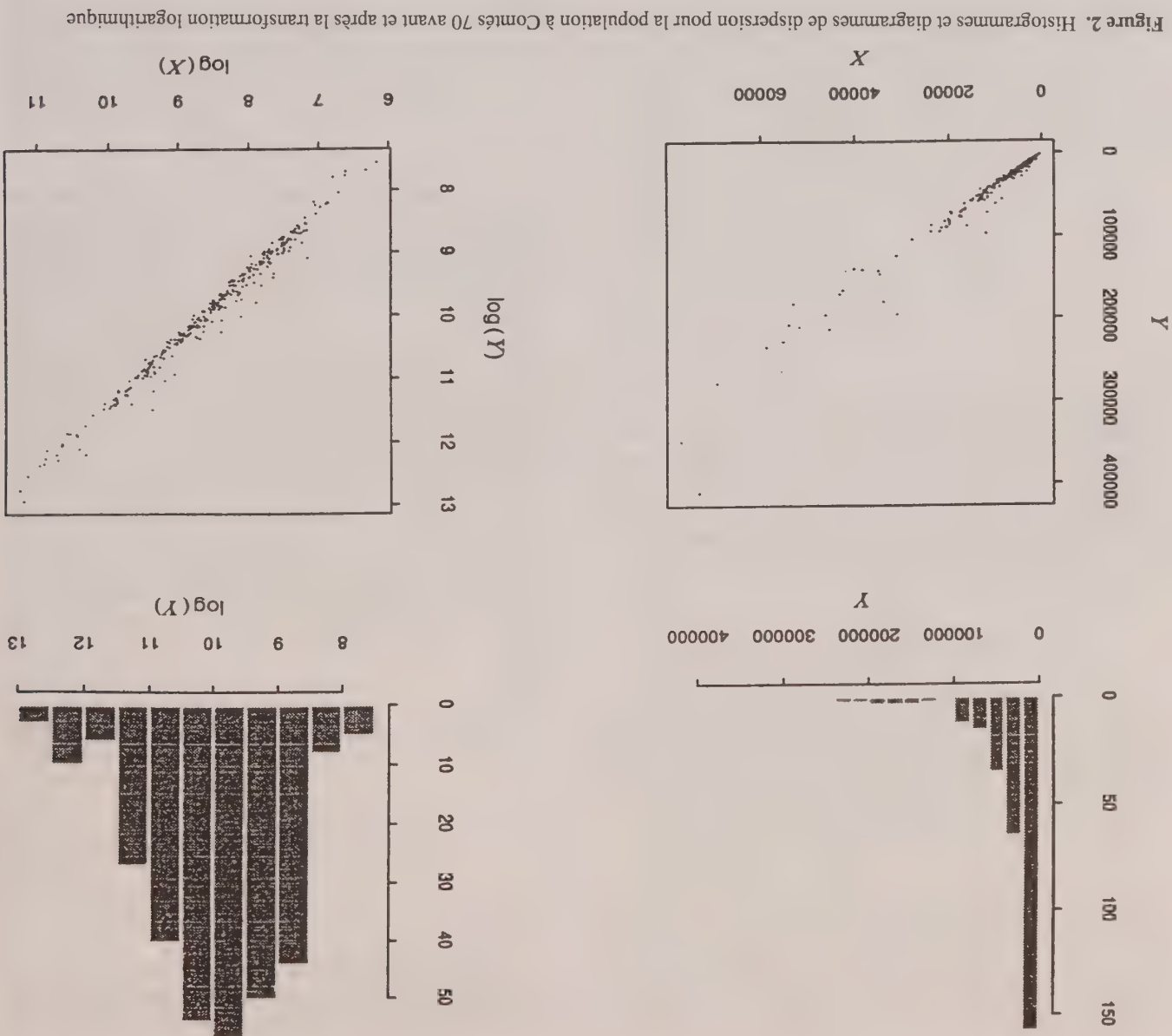


Figure 3. Graphiques des taux de non-couverture conditionnelle pour la population Comités 70 à partir de 10,000 échantillons de taille 32. Les lignes de référence ont pour valeur 2.5% et le taux attendu de non-couverture s'établit à 5%.

Ventes, la nouvelle méthode produit des résultats très moyens, vraisemblablement parce que la transformation logarithmique (ou toute transformation exponentielle) affaiblit effectivement la relation linéaire qui existe entre  $x$  et  $y$ . Nous avons également effectué des simulations pour des tailles d'échantillon égales à 16 et à 64, et (ou) pour un taux de couverture cible de 90%. Les résultats sont fort semblables à ceux que nous venons de présenter.

de  $\hat{y}$ ; les deux graphiques du milieu montrent les taux de non-couverture correspondant à notre nouvelle méthode. Le graphique qui figure en bas à gauche sera expliqué à la section 4. On voit clairement que notre nouvelle méthode avec transformation logarithmique produit une amélioration importante. Elle produit aussi une certaine amélioration pour les populations Villes, Comités 60 et Hôpitaux (les graphiques ne sont pas présentés ici). Pour les populations Cancer et



est égale à 32, et nous calculons  $\bar{x}_s$ ,  $\bar{y}$ ,  $\hat{\alpha}$ ,  $\hat{\beta}$  et construisons un intervalle de confiance de 95%  $I_{32}^*$ . Nous répétons le processus 10,000 fois pour chaque population. Les résultats sont présentés au tableau 2 sous le titre «Méthode de transformation» quand on connaît toutes les valeurs  $x_i$ , et sous le titre «Méthode de la vraisemblance empirique» quand on ne connaît que  $\bar{x}$ . Le terme «ratio» représente, pour chaque population, la longueur moyenne de l'intervalle de confiance divisée par la racine carrée de l'erreur quadratique moyenne. Le taux de non-couverture (Ncr) correspond à la proportion d'intervalle qui ne contiennent pas la moyenne de la population  $\bar{y}$ . Les chiffres qui figurent sous le titre «Méthode de régression (variance jackknife)» et «Méthode de régression (variance jackknife)» sont obtenus par la même méthode que celle de Royall and Cumberland (1981b) quand on utilise les variances de régression et jackknife habituelles de  $\bar{y}$ , respectivement, mais pour 10,000 échantillons aléatoires au

Hier des 10,000 échantillons originaux. Les résultats donnés sous le titre « Méthode de vraisemblance empirique (population créée) » seront expliqués à la section suivante.

Puis, à l'instar de Royall et Cumberland, nous faisons une inférence basée sur le plan de sondage et nous étudions les propriétés conditionnelles de couverture pour plusieurs méthodes d'estimation de l'intervalle. Plus précisément, nous répartissons les intervalles de confiance en 20 groupes d'après la taille de  $\bar{x}_j$  et nous représentons graphiquement les proportions d'intervalles qui, dans chaque groupe, ne contiennent pas la moyenne de la population  $\bar{y}$ . Pour chaque groupe, la proportion d'intervalles situés au-dessus (en-dessous) de  $\bar{y}$  est inscrite au-dessus (en-dessous) de la ligne horizontale. La figure 3 contient les graphiques obtenus pour la population Comités 70. Les deux graphiques supérieurs montrent les taux de non-couverture pour la méthode de régression fondée sur la variance de régression habituelle et sur la variance jackknife



estimateurs sont asymptotiquement plus efficaces quand on impose des contraintes que quand on n'en impose pas. Plus précisément, nous estimons  $F_N(u)$  dans (2.4) par

$$F_N(u) = \sum_{i \in s} p_i I[u_i \leq u], \quad (2.8)$$

où les  $p_i$  sont choisis en maximisant l'expression

$$p_i \geq 0, \quad \sum_{i \in s} p_i = 1, \quad \sum_{i \in s} p_i x_i = \bar{x}. \quad (2.10)$$

Si nous considérons que les  $y_i$ ,  $i \in s$  sont des réalisations des variables aléatoires  $Y_i$ ,  $i \in s$ , avec la fonction de distribution  $F$ , nous pouvons définir les  $p_i$  qui figurent dans (2.9) par  $p_i = F(Y_i) - F(Y_{i-1})$ . Owen (1990) donne à (2.9) le nom de fonction empirique de vraisemblance.

Deville et Särndal (1992) examinent la méthode susmentionnée sous l'angle de calage. Ils proposent d'utiliser des poids inégaux pour différentes unités tirées dans l'échantillon de façon à refléter la diversité de leur contribution, tout en maintenant  $\sum p_i x_i = \bar{x}$ . On estime que si ces poids donnent une estimation parfaite de  $\bar{x}$ , ils devraient également convenir pour l'estimation de  $\bar{y}$ .

Les expressions de (2.9) et (2.10) n'ont pas de solution si la valeur minimale de  $x$  dans un échantillon est égale ou supérieure à  $\bar{x}$  ou si la valeur maximale de  $x$  dans un échantillon est égale ou inférieure à  $\bar{x}$ . Le cas échéant, on peut pallier le problème en remplaçant (2.9) par

$$\sum_{i \in s} (np_i - 1)^2, \quad (2.11)$$
$$\sum_{i \in s} p_i = 1, \quad \sum_{i \in s} p_i x_i = \bar{x}. \quad (2.12)$$

subordonnée à une contrainte moins rigoureuse

En vertu de (2.11) et de (2.12), nous pouvons écrire

$$p_i = \frac{1}{n} + (\bar{x} - x_s)(x_i - \bar{x}) / \sum_{i \in s} (x_i - \bar{x})^2, \quad (2.13)$$

qui existe toujours à moins que tous les  $x_i$  de l'échantillon soient identiques. Cette dernière situation correspond à l'absence d'une covariable, ce qui implique que  $p_i = n^{-1}$  si  $\bar{x} = x_i$ , ou que la solution n'existe pas si  $\bar{x} \neq x_i$ . La fonction donnée en (2.11), qu'on dénomme vraisemblance euclidienne, est asymptotiquement équivalente à la vraisemblance empirique (2.9) (Owen 1990).

Pour l'étude de simulation que nous présentons à la section 3, nous proposons d'utiliser une correction du biais dans les calculs. Si  $h(w) = g(w) = \log(w)$ , nous proposons d'utiliser pour  $\bar{y}$  l'estimateur corrigé

$$\hat{y}^*(\alpha, \beta) = \int_{-\infty}^{\infty} \exp \left\{ \alpha + \beta u_i + \frac{1}{2} \sigma^2 \right\} F_N(u) du, \quad (2.14)$$

si on connaît tous les  $u_i$ ,  $i = 1, \dots, N$ , et de remplacer  $F_N(u)$  par  $F_N^N(u)$  et  $\bar{u}_N$  qui figure dans (2.6) par  $\bar{u}_s$  si on ne connaît que  $\bar{x}$ . Cette correction est motivée par des considérations relatives à la modélisation quand on émet l'hypothèse que la distribution est normale. De la même façon, on corrige  $I_n$  dans (2.7) comme suit

$$I_n^* = \{\hat{y}^*(\alpha, \beta) : (\alpha, \beta) \in C_n\}. \quad (2.15)$$

Si on effectue d'autres transformations exponentielles, on peut apporter des corrections similaires en s'appuyant sur les résultats de Pankratz et Dudley (1987).

### 3. APPLICATION À SIX POPULATIONS RÉELLES

Les six populations réelles étudiées par Royall et Cumberland (1981a, 1981b, 1985) sont décrites sommairement dans le tableau 1. Le choix de ces populations a été guidé par le souci d'obtenir divers types de données (démographiques, économiques, etc.) et diverses relations logiques entre les variables  $x$  et  $y$ . Il convient de souligner que nous avons ajouté une unité aux valeurs  $y$  de la population intitulée «Cancer» pour pouvoir appliquer la transformation logarithmique.

Tableau 1  
Résumés pour les six populations

Population	$N$	$\bar{x}$	$\bar{y}$	$\rho(x, y)$	$\rho(\log(x), \log(y))$
Cancer	301	$1.1288 \times 10^4$	$4.0847 \times 10^1$	0.967	0.948
Villes	125	$2.6602 \times 10^5$	$2.8553 \times 10^5$	0.947	0.953
Comtés 60	304	$8.9312 \times 10^3$	$3.2916 \times 10^4$	0.998	0.998
Comtés 70	304	$8.9312 \times 10^3$	$3.6984 \times 10^4$	0.982	0.991
Hôpitaux	393	$2.7470 \times 10^2$	$8.1465 \times 10^2$	0.911	0.943
Ventes	331	$2.3164 \times 10^9$	$2.4078 \times 10^9$	0.997	0.985

Les données correspondant à Comtés 70 sont représentées graphiquement à la figure 2. L'histogramme de  $y$  indique clairement que la distribution de la population est fortement asymétrique, tandis que le tracé de  $\log(y)$  témoigne d'une amélioration considérable. En outre, le diagramme de dispersion de  $\log(y)$  en fonction de  $\log(x)$  fait apparaître une meilleure relation linéaire que le diagramme de dispersion de  $y$  en fonction de  $x$ . La nécessité et l'avantage de la transformation sont donc évidents. Des observations similaires peuvent aussi être faites pour les populations Villes, Comtés 60 et Hôpitaux. Pour les populations Cancer et Ventes, la transformation logarithmique (ou toute autre transformation exponentielle) semble affaiblir la relation linéaire qui existe entre  $x$  et  $y$ .

Puis, nous illustrons notre nouvelle méthode en supposant que  $h = g = \log$  dans (2.2). Nous utilisons les équations (2.9) à (2.15) pour effectuer les calculs. Comme l'ont fait Royall et Cumberland (1981b, 1985), nous prenons pour chacune des six populations un échantillon aléatoire simple dont la taille

particulière dans notre étude de simulation. Des corrections générales sont à l'étude.

D'après le théorème 2.1,  $G_n$  est un intervalle de confiance

prudent pour  $\bar{y}(\alpha_N, \beta_N)$  qu'on peut aussi considérer comme un intervalle de confiance approximatif pour  $\bar{y}$ . Pour

améliorer le taux de couverture de  $G_n$  on observera que, dans un voisinage restreint de  $O = (\alpha, \beta)$ , les contours de  $\hat{y}(\alpha, \beta)$

sont approximativement des lignes droites parallèles dans le plan  $\alpha\beta$  (voir la figure 1). Soit  $(a, b)$  les cosinus directionnels

de la direction  $\overrightarrow{EF}$  le long de laquelle les contours augmentent. Alors,  $\hat{y}(\alpha, \beta)$  est (approximativement) une

fonction monotone de  $T_n = a(\alpha - \alpha) + b(\beta - \beta)$ , où  $T_n$  représente la variation le long de la direction  $\overrightarrow{EF}$  correspondant

aux variations de  $\alpha$  et  $\beta$ . Un choix naturel pour exprimer  $B_n$  est

$$B_n = \{(\alpha, \beta) : |a(\alpha - \alpha) + b(\beta - \beta)| \leq c \cot(\gamma/2; n - 2)\},$$

où  $c^2 = \text{Var}(T_n)/\sigma^2$ ,  $\text{Var}(T_n)$  est la variance de  $T_n$ , et

distribution de  $t$  pour  $n - 2$  degrés de liberté. Cette expression de  $B_n$  correspond à la région comprise entre les deux lignes

droites parallèles  $AB$  et  $CD$  de la figure 1.

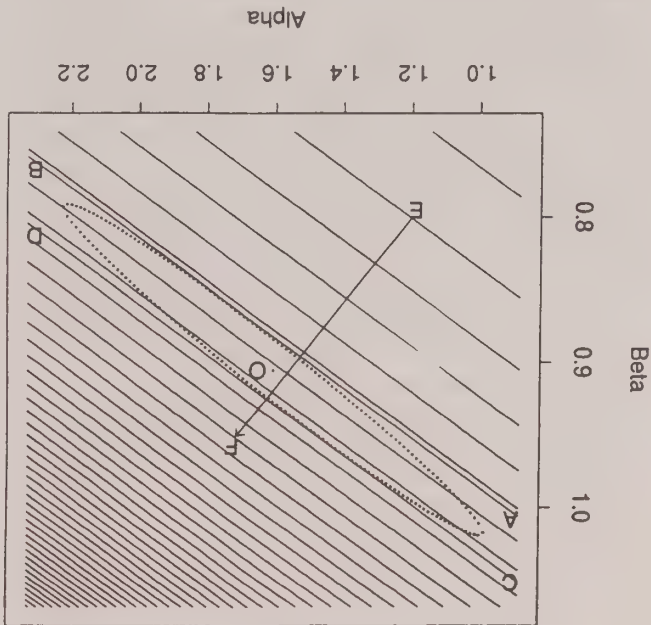


Figure 1. Contours de la fonction bivariable  $\hat{y}(\alpha, \beta)$  dans le voisinage de  $O = (\alpha, \beta)$ , basé sur un échantillon aléatoire de 32 unités de la population Cancer

Un inconvénient de l'expression de  $B_n$  susmentionnée tient à ce qu'elle représente une région *sans bornes* et que si les contours de  $\hat{y}(\alpha, \beta)$  ne sont pas pratiquement parallèles et (ou) droits, elle donne des intervalles de confiance très modérés. Pour empêcher que cela ne se produise, nous construisons une région elliptique bornée  $C_n$ , définie par les valeurs de  $(\alpha, \beta)$  qui satisfont

$$\left\{ n(\alpha - \alpha)^2 + 2n\alpha_s(\alpha - \alpha)(\beta - \beta) + \left( n\left( \frac{\bar{u}_s}{n^2} + r_s \right)^{-1} \sum_{i=1}^{i=s} (u_i - \bar{u}_s)^2 \right) (\beta - \beta)^2 \right\} \leq \left( 1 - \frac{N}{n} \right) \theta^2 t^2(\gamma/2; n - 2),$$

où  $(1 - n/N)$  fait partie de la variance de  $\alpha$  et  $\beta$ , car nous effectuons l'échantillonnage sans remise d'une population finie et

$$r_s = \frac{n^{-1} \sum_{i=1}^{i=s} (u_i - \bar{u}_N)^2 (v_i - \alpha - \beta u_i)^2}{n^{-1} \sum_{i=1}^{i=s} (u_i - \bar{u}_N)^2 (v_i - \alpha - \beta u_i)^2}.$$

(2.6)

est une estimation par sondage de la quantité  $r$  du théorème 2.1. Le  $C_n$  défini ainsi est représenté par la région comprise à l'intérieur de l'ellipse de la figure 1 et a la propriété de toucher les deux lignes de délimitation de  $B_n$  indépendamment de la direction  $(a, b)$ . Par conséquent, quand  $\hat{y}(\alpha, \beta)$  est effectivement une fonction monotone de  $T_n$ ,  $C_n$  produit pour  $\bar{y}$  le même intervalle de confiance que  $B_n$ . Cependant,  $C_n$  est moins vulnérable que  $B_n$  si les contours de  $\hat{y}(\alpha, \beta)$  ne sont pas pratiquement parallèles et (ou) droits, car  $C_n$  tend vers un point à mesure que  $n$  augmente. L'intervalle de confiance de  $\bar{y}$  correspondant à  $C_n$  est défini comme

$$I_n = \{ \hat{y}(\alpha, \beta) : (\alpha, \beta) \in C_n \}. \quad (2.7)$$

Les distributions de l'erreur étant plus symétriques après la transformation, le nouvel intervalle de confiance fondé sur  $C_n$  devrait être meilleur que celui construit sans transformation. On notera que, puisque tous les  $x_i$  sont connus, d'autres méthodes, dont la stratification optimale et la stratification postérieure, pourraient être meilleures. Cependant, la stratification optimale peut s'avérer impossible dans certains cas, comme l'explique Cochran (1977, p. 134). Il est également nécessaire d'étudier l'utilisation de la stratification à posteriori quand les distributions d'erreur sont fortement asymétriques. Nous passons maintenant à la discussion du deuxième cas où on connaît  $\bar{x} = (x_1 + \dots + x_N)/N$ , mais où on ne connaît pas  $x_i$ ,  $i = 1, \dots, N$ . Pour procéder de la même façon que dans le premier cas, une méthode consiste à estimer  $F_N(u)$  et à trouver une façon d'utiliser l'information que donne  $\bar{x}$ . Nous constatons que la méthode de vraisemblance empirique qui suit est un moyen efficace de le faire. Nous nous limitons ici à la description des notions principales; le lecteur que la question intéresse devrait consulter Owen (1988, 1990) et Chen et Qin (1992) pour plus de détails. L'idée principale consiste à maximiser les fonctions (empiriques) de vraisemblance pour diverses contraintes formulées d'après ce qu'on sait de certains aspects des paramètres. Par exemple, dans notre cas, nous connaissons  $\bar{x}$ . Chen et Qin (1992) ont montré que les



Il convient de souligner que les modèles (2.1) et (2.2) sont utilisés ici pour amener les transformations, les estimateurs ponctuels ou les calculs d'intervalle de confiance. Cependant, nous fonderons notre étude de taux conditionnels de couverture sur la mesure de probabilité générée par le plan de sondage, comme le font Royall et Cumberland (1985). À cette fin, nous insérons notre population finie dans une série de populations portant l'indice  $k$ . Autrement dit, nous devrions utiliser un indice inférieur  $k$  pour écrire  $N = N_k$  et  $n = n_k$ , etc. Toutefois, pour des raisons de simplicité, nous supprimons l'indice  $k$  quand cela ne pose aucun risque de confusion. Soit  $v_i = h(y_i)$ ,  $u_i = g(x_i)$ ,  $v_N = N^{-1} \sum_{i=1}^N v_i$  et  $u_N = N^{-1} \sum_{i=1}^N u_i$ . Définissons

$$\beta_N = \frac{\sum_{i=1}^N (u_i - \bar{u}_N) v_i}{\sum_{i=1}^N (u_i - \bar{u}_N)^2},$$

$$\alpha_N = \bar{v}_N - \beta_N \bar{u}_N,$$

$$e_i = v_i - (\alpha_N + \beta_N u_i),$$

$$\sigma_N^2 = \frac{1}{N-1} \sum_{i=1}^N e_i^2.$$

Supposons que  $s \subset S$  est un échantillon aléatoire simple de taille  $n$ . Nous définissons de façon similaire

$$\hat{\beta} = \frac{\sum_{i \in s} (u_i - \bar{u}_s) v_i}{\sum_{i \in s} (u_i - \bar{u}_s)^2},$$

$$\hat{\alpha} = \bar{v}_s - \hat{\beta} \bar{u}_s,$$

$$\sigma^2 = \frac{1}{2} \sum_{i \in s} (v_i - \hat{\alpha} - \hat{\beta} u_i)^2,$$

où  $\bar{u}_s$  et  $\bar{v}_s$  sont les moyennes d'échantillon.

Représentons la fonction inverse de  $h(\cdot)$  par  $h^{-1}(\cdot)$ . Alors, la valeur ajustée de  $y_i$  est

$$\hat{y}_i = h^{-1}(\hat{\alpha} + \hat{\beta} u_i). \quad (2.3)$$

Nous discutons de l'estimation de l'intervalle de confiance de  $\bar{y}$  dans deux cas. Dans le premier, où tous les  $x_i$  ( $i = 1, \dots, N$ ) sont connus,  $(\sum_{i \in s} y_i + \sum_{i \notin s} \hat{y}_i)/N$  est un estimateur naturel de  $\bar{y}$ . Cependant, en vue de construire les intervalles de confiance de  $\bar{y}$ , nous étudions plutôt la distribution de

$$\hat{y}(\hat{\alpha}, \hat{\beta}) = \frac{1}{N} \sum_{i=1}^N \hat{y}_i = \int_{-\infty}^{\infty} h^{-1}(\hat{\alpha} + \hat{\beta} u) dF_N(u) \quad (2.4)$$

où  $F_N(u)$  est la fonction empirique de distribution de  $u_i$  ( $i = 1, \dots, N$ ). Manifestement, la distribution de  $\hat{y}(\hat{\alpha}, \hat{\beta})$  est

déterminée par la distribution de  $(\hat{\alpha}, \hat{\beta})$  qui est décrite dans le théorème axé sur le plan de sondage qui suit.

**Théorème 2.1** Supposons que quand  $k \rightarrow \infty$ ,  $n = n_k$  et  $N - n = N_k - n_k$  tendent aussi vers  $\infty$  et que

$$1. \bar{u} = \lim_{k \rightarrow \infty} N^{-1} \sum_{i=1}^N u_i \text{ existe.}$$

$$2. N^{-1} \sum_{i=1}^N u_i^4 = O(1).$$

$$3. \sigma_u^2 = \lim_{k \rightarrow \infty} \sigma_{u_N}^2 = \lim_{k \rightarrow \infty} (N-1)^{-1} \sum_{i=1}^N (u_i - \bar{u}_N)^2 \text{ existe et est plus grande que zéro.}$$

$$4. \sigma^2 = \lim_{k \rightarrow \infty} \sigma_N^2 = \lim_{k \rightarrow \infty} (N-1)^{-1} \sum_{i=1}^N e_i^2 \text{ existe et est plus grande que zéro.}$$

$$5. N^{-1} \sum_{i=1}^N |e_i|^3 = O(1), \quad N^{-1} \sum_{i=1}^N (u_i - \bar{u}_N) e_i^3 = O(1).$$

$$6. r = \lim_{k \rightarrow \infty} (\sigma_{u_N}^2)^{-1} N^{-1} \sum_{i=1}^N (u_i - \bar{u}_N)^2 e_i^2 \text{ existe et est plus grande que zéro.}$$

$$7. f = \lim_{k \rightarrow \infty} n/N \text{ existe et est plus petit que 1.}$$

Alors,

$$(1) \text{ la distribution de } \sqrt{n}(\hat{\alpha} - \alpha_N, \hat{\beta} - \beta_N)' \text{ converge vers la distribution normale à deux variables } N_2(0, \Sigma), \text{ où}$$

$$\Sigma = \begin{pmatrix} 1 + \frac{u^2}{r} - \frac{\sigma_u^2}{r} & -\frac{\bar{u}}{r} - \frac{\sigma_{u^2}}{r} \\ -\frac{\bar{u}}{r} - \frac{\sigma_{u^2}}{r} & \frac{1}{r} - \frac{\sigma_u^2}{r} \end{pmatrix} (1-f)\sigma^2.$$

(2) Posons que  $B_n$  représente n'importe quelle région de confiance 100(1 -  $\gamma$ )% commune pour  $(\alpha_N, \beta_N)$  et définissons  $G_n$  par

$$G_n = \{ \hat{y}(\hat{\alpha}, \hat{\beta}) : (\hat{\alpha}, \hat{\beta}) \in B_n \}. \quad (2.5)$$

Alors,

$$\text{Prob} \{ \bar{y}(\alpha_N, \beta_N) \in G_n \} \geq 1 - \gamma,$$

$$\text{où } \bar{y}(\alpha_N, \beta_N) = \sum_{i=1}^N h^{-1}(\alpha_N + \beta_N u_i)/N.$$

La preuve est exposée en annexe.

Nous notons que, sans distribution normale sous-jacente des erreurs, il n'est pas facile d'obtenir une région de confiance  $B_n$  exacte pour  $(\alpha_N, \beta_N)$  pour un niveau de confiance particulier  $1 - \gamma$ . La région de confiance  $B_n$  utilisée dans la discussion qui suit et les expressions qui en découlent sont donc approximatives.

Le théorème 2.1 nous permet de construire des intervalles de confiance pour  $\bar{y}(\alpha_N, \beta_N)$ , mais  $\bar{y}(\alpha_N, \beta_N)$  n'est généralement pas égal à  $\bar{y}$ . Il s'agit d'un problème intrinsèque à condition d'utiliser une transformation non linéaire. Dans les cas où seul un estimateur ponctuel est nécessaire, nous recommandons d'utiliser l'estimateur de régression en ce moment

et de suivre la méthode élaborée dans le présent article pour estimer l'intervalle. Toutefois, il est possible d'apporter des corrections de biais à  $\hat{y}(\hat{\alpha}, \hat{\beta})$  et nous utilisons une correction

obtenus lors de l'étude d'une autre méthode de construction des intervalles de confiance et de son application aux six populations étudiées par Royall et Cumberland et par de nombreux autres chercheurs. Comme nous le montrerons à la section 3, le taux conditionnel de couverture de nos intervalles de confiance est plus exact.

Nous appliquons concomitamment deux concepts importants, nommément la transformation et la vraisemblance empirique, pour tenter de résoudre les problèmes auxquels se sont heurtés Royall et Cumberland en particulier, et pour mettre au point une nouvelle méthode générale. Comme l'explique Cochran (1977, p. 150), les spécialistes de la théorie des enquêtes par sondage préfèrent faire, au plus, des suppositions limitées quant à la distribution de fréquences des données d'échantillon. Cependant, l'utilisation de l'estimateur de ratio ou de l'estimateur de régression peut augmenter l'exactitude, car ces estimateurs tiennent compte de la corrélation de  $y_i$  et  $x_i$ . On peut évidemment décrire cette corrélation en émettant certaines hypothèses, comme l'existence d'une relation linéaire approximative entre  $y$  et  $x$ . Bien que pratiquement aucune autre hypothèse ne soit nécessaire pour utiliser la méthode du ratio ou la méthode de régression, l'équation (1.1) se fonde manifestement sur une approximation normale. Or, il est bien connu que l'approximation normale donne des résultats très médiocres si la distribution de la population est fortement asymétrique et que l'effectif de l'échantillon est petit. En ce qui concerne la méthode (1.1), la construction des intervalles de confiance est d'autant meilleure que la distribution de l'estimateur est proche de la distribution normale. Si la distribution de la population est fortement asymétrique, on peut effectuer une transformation en vue de produire une distribution de population qui soit, au moins, plus symétrique, de sorte que l'approximation normale faite pour l'estimateur soit plus exacte.

Quand on utilise les estimateurs de ratio et de régression, il est essentiel de connaître  $\bar{x}$  pour obtenir de meilleurs résultats que ceux donnés par la moyenne de l'échantillon. La méthode que nous proposons permet d'intégrer toutes les données dont on dispose au sujet de la variable auxiliaire  $x$ . Toutefois, si  $\bar{x}$  est la seule donnée auxiliaire disponible, il est difficile de l'utiliser directement quand une transformation est nécessaire, car toute transformation non linéaire obscurcit le lien entre  $x$  et  $y$ . Le cas échéant, nous constatons que la méthode de vraisemblance empirique est un moyen utile de résoudre le problème; à cet égard, on consultera notamment Owen (1988, 1990) et Chen et Qin (1992). Dans ces conditions, la méthode de la vraisemblance empirique peut aussi être considérée comme une méthode de calage, ainsi que l'expose Deville et Särndal (1992). Cette méthode nous permet de ne pas perdre d'information au sujet de  $x$  après avoir transformé les données.

De nombreux auteurs ont examiné la façon d'utiliser les transformations pour améliorer les inférences concernant les échelles transformées (Box et Cox 1964; Carroll et Rupert 1988; Calvin et Sedransk 1991, ainsi que les auteurs qu'ils mentionnent). D'autres ont aussi étudié les moyens de faire des inférences au sujet de l'échelle originale après avoir effectué une transformation (Carroll et Rupert 1984; Elliott

1977). La nouveauté de notre méthode tient à ce que qu'elle vise à relier les deux étapes susmentionnées grâce à la combinaison de la transformation et des données auxiliaires et (ou) à l'application d'une méthode de vraisemblance empirique, au besoin.

Nous décrivons notre méthode en détail à la section 2. Puis, à la section 3, nous l'appliquons aux six populations étudiées par Royall et Cumberland. À la section 4, nous démontrons la validité de notre méthode dans des conditions arbitraires et nous concluons l'article par certains commentaires.

## 2. LA NOUVELLE MÉTHODE

Comme nous l'avons mentionné à la section précédente, un des problèmes que pose l'intervalle de confiance (1.1) tient à ce qu'il donne des résultats incorrects si la distribution de  $(\bar{y} - \bar{y})/\sqrt{V}$  est fortement asymétrique et très différente de la distribution normale. Le problème peut découler de l'asymétrie de la distribution de la population. Si l'asymétrie est forte, l'application d'un intervalle de confiance central tel que (1.1) est vouée à l'échec. Le modèle de base employé par Royall et Cumberland (1981a, 1981b, 1985) est

$$(2.1) \quad y_i = \alpha + \beta x_i + \epsilon_i,$$

où  $E(\epsilon_i) = 0$ ,  $V(\epsilon_i) = \sigma^2$  et  $\text{Cov}(\epsilon_i, \epsilon_j) = 0$ , pour  $i \neq j$ . Il est facile de montrer que les distributions de l'erreur correspondent aux six populations réelles étudiées par Royall et Cumberland sont très asymétriques. Ces observations nous mènent à considérer la transformation des variables  $y$  et (ou)  $x$ , et l'application du modèle

$$(2.2) \quad h(y_i) = \alpha + \beta g(x_i) + \sigma \epsilon_i,$$

où  $h(\cdot)$  et  $g(\cdot)$  sont deux fonctions monotones. Nombre de familles de transformation sont proposées dans la littérature. Une de celles utilisées couramment est la transformée exponentielle de Box-Cox définie par

$$f(x, \lambda) = \begin{cases} (x^\lambda - 1)/\lambda & \text{quand } \lambda \neq 0, \\ \log(x) & \text{quand } \lambda = 0. \end{cases}$$

Le modèle (2.1) est un cas spécial de (2.2) dans lequel les fonctions  $h$  et  $g$  sont toutes deux égales à  $f(x, 1)$ . En ce qui concerne le modèle (2.2), le choix des transformations pourrait se fonder sur l'examen des valeurs de  $x$  et d' $y$  dans l'échantillon afin de déterminer s'il existe une relation modélisable entre ces variables, ou sur des connaissances particulières au sujet de la population étudiée. Par exemple, dans le cas des six populations examinées par Royall et Cumberland, les distributions de population sont fortement étalées vers la droite, caractéristique que l'on peut déduire de la nature finie de ces populations. Par conséquent, une transformation logarithmique pourrait les rendre moins asymétriques. Nous examinerons d'autres méthodes, plus objectives, de choix des transformations à la section 4.



# Une méthode de transformation applicable à l'échantillonnage de populations finies calée par une méthode de vraisemblance empirique

GEMAI CHEN et JIAHUA CHEN<sup>1</sup>

## RÉSUMÉ

Dans le présent article, nous présentons une méthode qui permet d'estimer l'intervalle de confiance de la moyenne d'une population finie quand on dispose de certaines données auxiliaires. Comme l'ont montré Royall et Cumberland grâce à une série d'études empiriques, l'application native des méthodes existantes de construction des intervalles de confiance de la moyenne d'une population aboutit parfois à de très médiocres probabilités conditionnelles de couverture subordonnées à la moyenne d'échantillon de la covariable. Le cas échéant, nous proposons de transformer les données pour améliorer la précision de l'approximation normale. Puis, d'après les données transformées, nous faisons une inférence quant à la moyenne de la population originale et intégrerons les données auxiliaires à l'inférence soit directement, soit par calage au moyen d'une fonction empirique de vraisemblance. Nous appliquons notre méthode, qui est basée sur le plan de sondage, à six populations réelles et constatons que, dans les cas où la transformation est nécessaire, elle donne de bons résultats comparativement à la méthode de régression habituelle.

**MOTS CLÉS :** Population finie; échantillonnage; intervalle de confiance; transformation; vraisemblance empirique.

## 1. INTRODUCTION

Supposons que  $(x_i, y_i), i = 1, 2, \dots, N$  représente les valeurs

associées à  $N$  unités d'une population finie. Pour l'unité  $i, y_i$  représente la variable étudiée et  $x_i$  la variable auxiliaire. Un des problèmes posés par les populations finies sur lesquels on s'est penché le plus est celui de l'estimation de la moyenne de la population  $\bar{y} = (y_1 + \dots + y_N)/N$  (ou du total  $N\bar{y}$ ) pour diverses méthodes d'échantillonnage. Nous nous concentrons ici sur l'échantillonnage aléatoire simple, parce que cette méthode est celle qui permet le mieux d'illustrer les problèmes que nous voulons étudier et qu'il est facile de généraliser les résultats obtenus à d'autres qui ont pour élément de base l'échantillonnage aléatoire simple.

Souvent, on dispose de renseignements sur la variable auxiliaire  $x$  qui permettent de faire des inférences au sujet de  $y$ . Par exemple, posons  $S = \{1, \dots, i, \dots, N\}$  et représentons par  $s \subset S$  un échantillon aléatoire simple de taille  $n$ . Quand la moyenne  $\bar{x} = (x_1 + \dots + x_n)/n$  est connue et que les variables  $x$  et  $y$  sont corrélées, on peut estimer la moyenne de la population  $\bar{y}$  grâce à l'estimateur de ratio  $\hat{\bar{y}} = (\bar{y}_s/\bar{x}_s)\bar{x}$  ou l'estimateur de régression  $\hat{\bar{y}} = \bar{y}_s + b(\bar{x} - \bar{x}_s)$ , où  $\bar{x}_s$  et  $\bar{y}_s$  représentent les moyennes de  $x$  et  $y$ , respectivement, et où  $b = \sum(x_i - \bar{x}_s)(y_i - \bar{y}_s) / \sum(x_i - \bar{x}_s)^2$ . Dans des conditions très générales, l'estimateur de ratio et l'estimateur de régression sont tous deux asymptotiquement normaux; à ce sujet, consulter Scott et Wu (1981), Bickel et Freedman (1984), et le théorème 2.1 à la section 2. Donc, si  $v$  est un estimateur choisi avec soin de la variance de  $\hat{\bar{y}}$ , on considère ordinairement que la variable standardisée  $(\hat{\bar{y}} - \bar{y})/\sqrt{v}$  obéit à la loi normale réduite. Par conséquent, si

$z_\alpha$  représente le percentile supérieur d'ordre  $\alpha$  de la

$$(1.1) \quad (\hat{\bar{y}} - \bar{y})/\sqrt{v} \approx z_\alpha \sqrt{v}, \quad \hat{\bar{y}} + z_\alpha \sqrt{v}$$

produira un intervalle de confiance approximatif de

$$100(1 - 2\alpha)\% \text{ pour } \bar{y}.$$

L'intervalle de confiance (1.1) est beaucoup utilisé dans la pratique. Cependant, son application à certaines populations pose des problèmes. Royall et Cumberland (1981a, 1981b, 1985) ont étudié les estimateurs de ratio et de régression, et les ont appliqués à six populations réelles pour lesquelles il semble exister de fortes corrélations entre  $x$  et  $y$ . (Consulter la section 3 pour une description sommaire des six populations.) Ces auteurs ont utilisé divers estimateurs de la variance de  $\hat{\bar{y}}$  et ont observé que le taux conditionnel actuel de couverture de l'intervalle de confiance (1.1), subordonné à  $\bar{x}_s$ , dépend fortement de la taille de  $\bar{x}_s$  et est généralement beaucoup plus faible que le taux de couverture proclamé, même dans le cas de l'estimateur de variance le plus souple. Par exemple, le taux conditionnel de couverture de l'intervalle de confiance de 95% pour la population qu'ils nomment Comités 70 se chiffre à 76% quand on applique l'estimateur jackknife de la variance et que la valeur de  $\bar{x}_s$  est faible, et peut avoir une valeur aussi faible que 50% quand on se sert d'autres estimateurs. Les études susmentionnées montrent qu'il est nécessaire de construire des intervalles de confiance «dignes de leur nom» (Royall et Cumberland 1985, p. 359). Cependant, jusqu'à présent, peu de progrès ont été accomplis dans ce domaine. Dans le présent article, nous présentons certains résultats

<sup>1</sup> Gemai Chen, Département de mathématique et de statistique, University of Regina, Regina, Saskatchewan, Canada, S4S 0A2; Jiahua Chen, Département de statistique et de sciences actuariales, University of Waterloo, Waterloo, Ontario, Canada, N2L 3G1.

- CLEVELAND, W.S. (1993). *Visualizing Data*. Summit, New Jersey: Hobart Press.
- DEVILLE, J.-C., et SÄRNDAAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- DEVILLE, J.-C., SÄRNDAAL, C.-E., et SAUTORY, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.
- ESTEVAO, V., HIDIROGLOU, M.A., et SÄRNDAAL, C.-E. (1995). Methodological principles for a generalized estimating system at Statistics Canada. *Journal of Official Statistics*, 11, 181-204.
- FULLER, W.A., LOUGHIN, M.M., et BAKER, H.D. (1993). Regression weighting for the 1987-1988, Nationwide Food Consumption Survey. Rapport soumis non-publié à United States Department of Agriculture.
- LEMAÎTRE, G., et DUFOUR, J. (1987). Une méthode intégrée de pondération des personnes et des familles. *Techniques d'enquête*, 13, 211-220.
- LOVE, S., ALEXANDER, C.H., et DALZELL, D. (1995). Constructing a major survey – operational plans and issues of continuous measurement. *Proceedings of the Section on Survey Research Methods, American Statistical Association*. À paraître.
- MCCARTHY, P.J. (1969). Pseudo-replication: half-samples. *Revue de l'Institut International de Statistique*, 37, 239-264.
- RAO, J.N.K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics*, 10, 153-165.
- SÄRNDAAL, C.-E., SWENSSON, B., et WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- STUKEL, D.M., et BOYER, R. (1992). Calibration estimation: an application to the Canadian Labor Force Survey. Documents de travail, SSMD-92-009E, Ottawa: Statistique Canada.
- ZIESCHANG, K.D. (1990). Sample weighting methods and estimation methods in the Consumer Expenditure Survey. *Journal of the American Statistical Association*, 85, 986-1001.



les données pour une année complète provenant des enquêtes par entrevue et journal, pour 1990. Les résultats ont été simplifiés à ceux que nous venons de présenter, et nous avons adopté un ensemble final de 24 variables auxiliaires, basé sur le nombre de personne par âge, race, sexe, région, urbain x région, sur le nombre d'UC par type de ménage et sur l'utilisation d'une coordonnée à l'origine. On procède actuellement à la conversion des estimations pour l'enquête CB, afin d'utiliser la régression restreinte pour celle-ci.

#### 4. CONCLUSION

L'objectif de cette étude était d'explorer des méthodes permettant de calculer les facteurs de pondération pour les ménages, facteurs qui ne dépendraient pas de la pondération accordée à un seul membre du ménage. Nous avons présenté différents types de facteurs de pondération basés sur l'estimation par régression, et nous avons évalué leurs mérites relatifs. L'estimation par régression incorpore les méthodes actuelles de stratification a posteriori de l'enquête, dans lesquelles la somme pondérée des personnes dans chaque strate a posteriori doit être égale aux chiffres correspondants fournis par un recensement indépendant. On y parvient en utilisant des variables auxiliaires qui sont incorporées dans le modèle de régression. Celui-ci produit également auto-matiquement, pour chaque ménage de l'échantillon, un facteur de pondération qui ne dépend d'aucun de ses membres.

Nous avons étudié huit types de facteurs de pondération, fournis par cinq modèles de régression différents. Afin d'éliminer les facteurs négatifs indésirables qui peuvent découler de toute procédure d'estimation par régression utilisant les moindres carrés ordinaires, nous avons adapté les estimations de régression restreinte au problème présent. La régression restreinte est justement assez souple pour restreindre les écarts possibles de chaque facteur de pondération final par rapport à son facteur de base, tout en respectant les propriétés mentionnées ci-dessus. On peut donc, notamment, imposer la contrainte que les facteurs soient positifs. Les facteurs de régression restreinte se calculent facilement au moyen de logiciels matriciels comme S-Plus<sup>TM</sup> ou SAS/IML<sup>TM</sup>.

La régression restreinte et, de façon plus générale, le calage restreint offrent un certain nombre d'attraits pour les enquêtes-ménages, comme celle que nous étudions ici, mais également pour les enquêtes portant sur d'autres types d'unités comme les hôpitaux, les écoles ou les établissements commerciaux pour lesquels on peut disposer d'un grand nombre de données auxiliaires. Compte tenu des données antérieures pour les variables cibles, on peut utiliser des procédures standard de construction de modèles afin de choisir les variables auxiliaires. Les propriétés de l'estimation par régression peuvent servir à choisir de façon optimale les variables prédictives, afin de réduire la redondance de l'information qui se trouve incorporée dans la procédure d'estimation pour l'enquête. C'est là un des plus grands avantages d'utiliser un estimateur ayant déjà fait l'objet de travaux nombreux et attestés. De bonnes variables prédictives peuvent comprendre des variables qualitatives, p. ex., âge, race, type d'hôpital (hôpital général, hôpital psychiatrique, etc.), type d'entreprise

(usine, vente au détail, etc.) et qui peuvent souvent être utilisées dans la stratification ordinaire ou a posteriori. Les variables prédictives peuvent également être des variables quantitatives comme le revenu familial, les ventes annuelles, le nombre d'étudiants de différents niveaux ou le nombre de journées d'hospitalisation, pour n'en nommer que quelques-unes. Dans notre application, le fait d'inclure une coordonnée à l'origine a également permis de réduire de manière notable les erreurs-types des estimations pour l'enquête. La méthode de régression permet également d'incorporer facilement, dans l'estimation, des données à des niveaux différents. Dans l'enquête-ménage étudiée ici, nous avons inclus des variables auxiliaires pour les personnes et les ménages.

La grande souplesse de la régression offre aux statisticiens des options qu'ils ne pourraient avoir autrement. Si de nouvelles variables prédictives pertinentes deviennent disponibles, les logiciels permettant de calculer les estimations par régression peuvent facilement les incorporer, il suffit alors de modifier la matrice des variables auxiliaires et les vecteurs des contrôles de population. Les logiciels écrits pour effectuer seulement la stratification a posteriori ou l'estimation des quotients avec une seule variable auxiliaire, par exemple, pourraient nécessiter des modifications importantes afin que l'on puisse changer l'estimateur. Il va de soi que si l'estimateur est l'un des types moins généraux, de quotient ou de stratification a posteriori, le logiciel de régression le traitera comme un cas spécial. Aux États-Unis, on envisage d'effectuer une très vaste enquête permanente sur les ménages (Love, Alexander et Dalzell 1995), qui fournirait des estimations très précises pour de nombreuses caractéristiques que l'on pourrait utiliser comme totaux de contrôle dans des enquêtes plus petites. La méthode de régression restreinte permettrait à l'enquête CB d'incorporer sans heurt ces nouvelles données dans les estimations, si elles étaient disponibles.

#### REMERCIEMENTS

Les opinions exprimées dans cet article sont celles des auteurs et ne représentent pas la politique du Bureau of Labour Statistic. Nous remercions les examinateurs et les rédacteurs pour leurs observations utiles.

#### BIBLIOGRAPHIE

- ALEXANDER, C.H. (1987). Une classe de méthodes utilisant des chiffres de population dans la pondération des ménages. *Techniques d'enquête*, 13, 193-209.
- BARDSLEY, P., et CHAMBERS, R.L. (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics*, 33, 290-299.
- BETHLEHEM, J.G., et KELLER, W.J. (1987). Linear weighting of sample survey data. *Journal of Official Statistics*, 3, 141-153.
- CASADY, R.J., et VALLANT, R. (1993). Propriétés conditionnelles des estimateurs de stratification a posteriori selon la théorie normale. *Techniques d'enquête*, 19, 193-203.

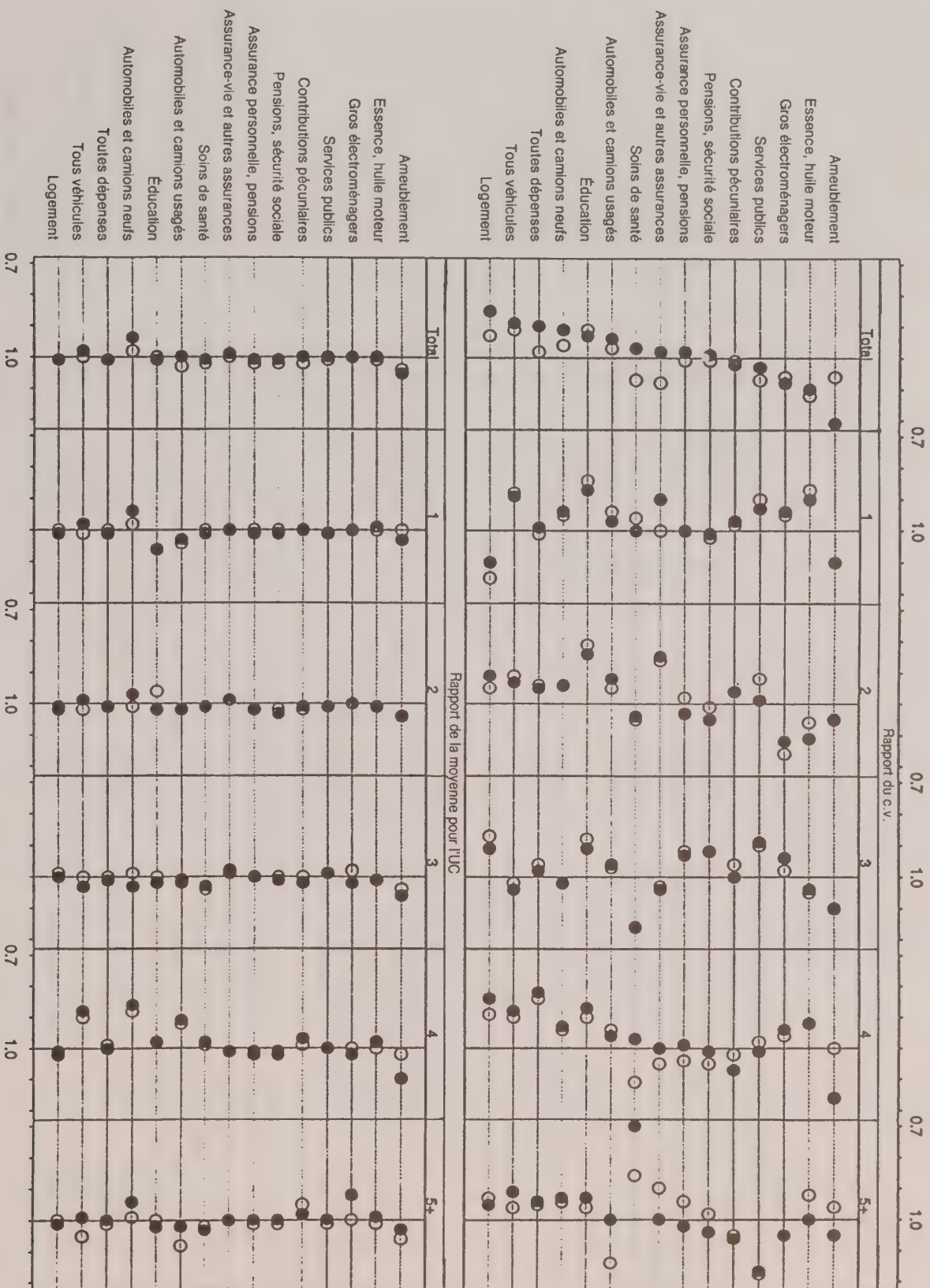
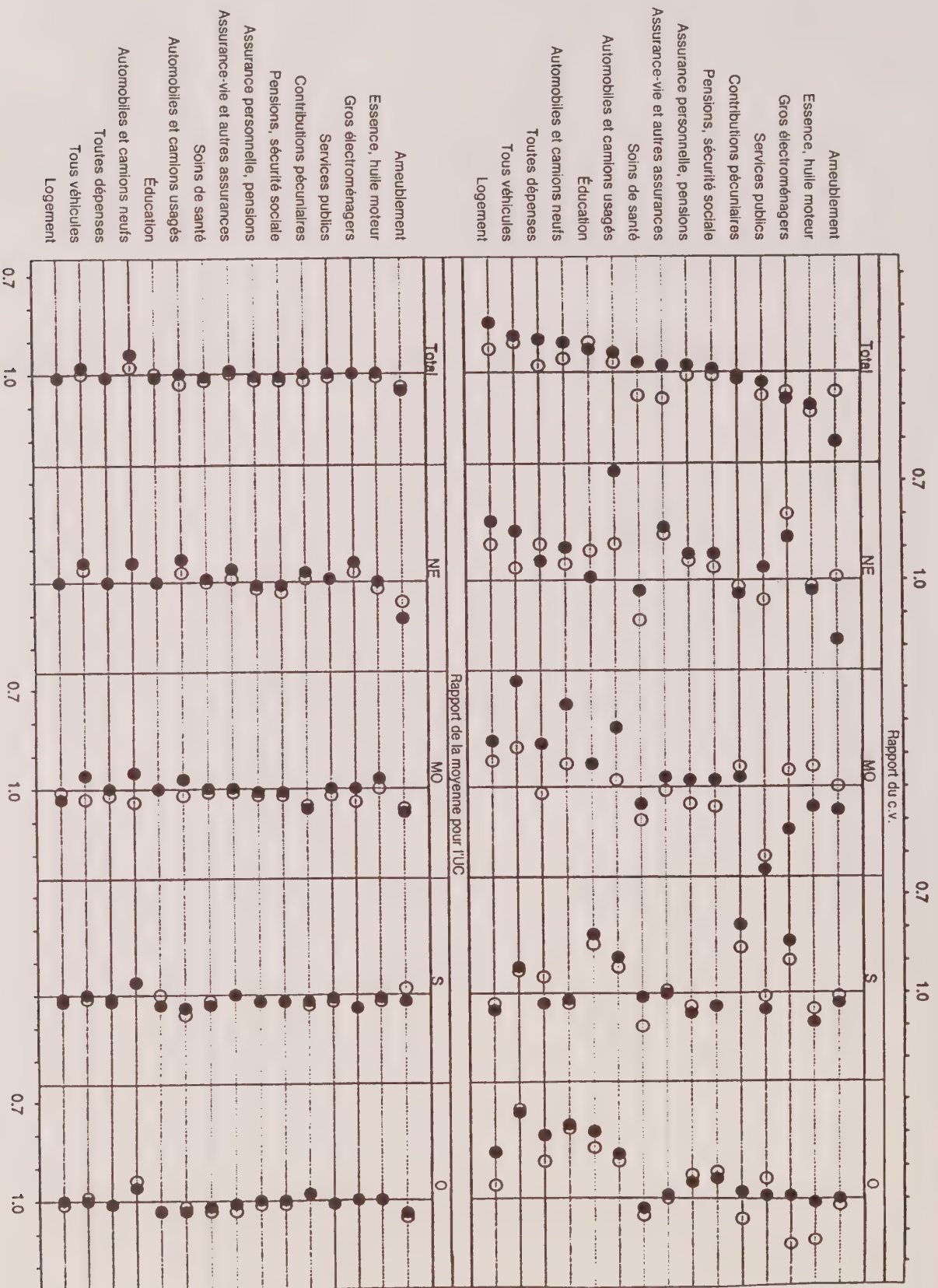


Figure 4. Rapports des c.v. et des moyennes obtenus par deux méthodes de pondération aux c.v. de la méthode PP, selon la taille de l'UC



Figure 3. Rapports des c.v. et des moyennes obtenus par deux méthodes de pondération aux c.v. de la méthode PP, selon la région

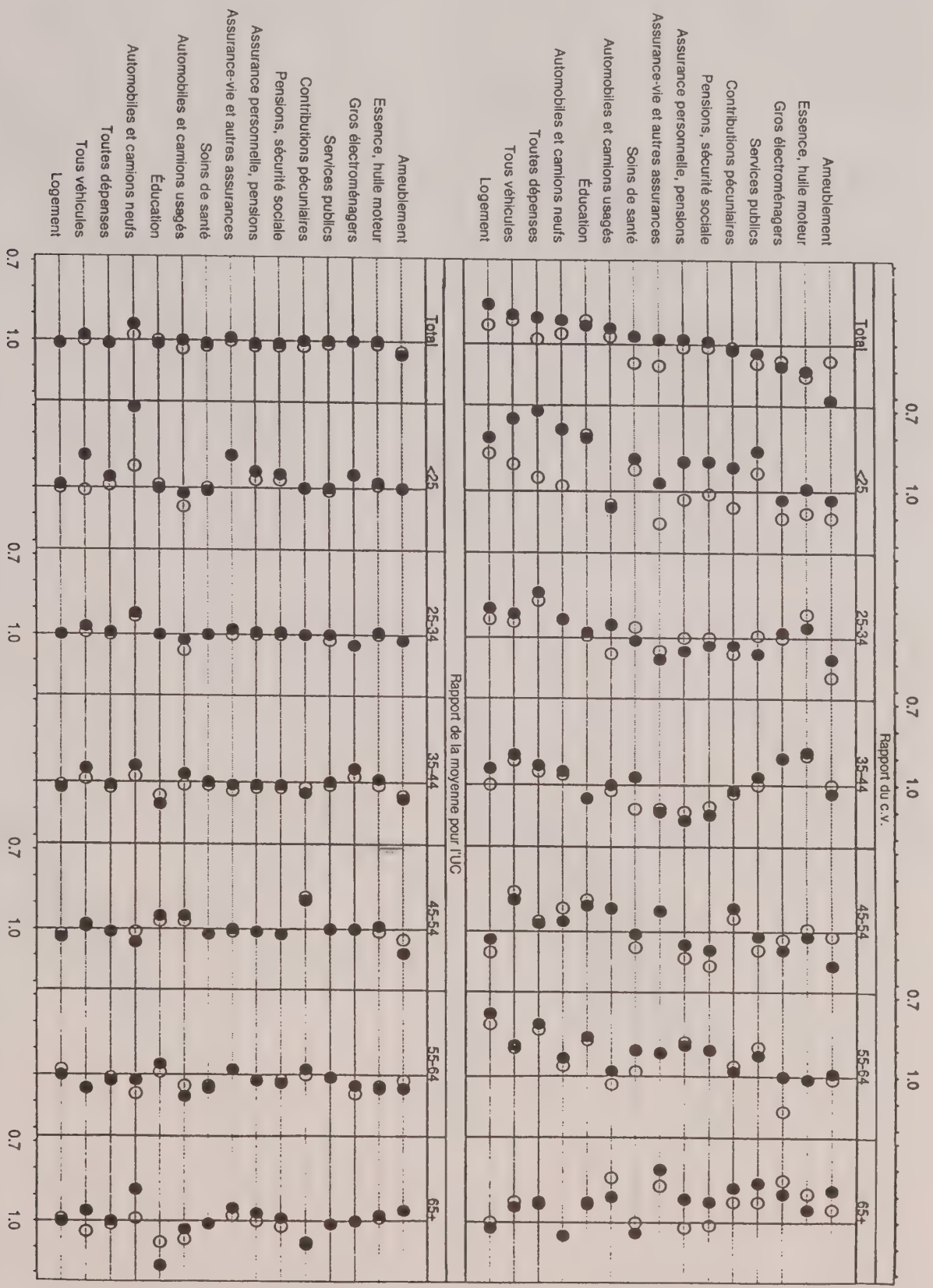


Rapports aux c.v. de la méthode PP

○ calwts0

● calwts1

Figure 2. Rapports des c.v. et des moyennes obtenus pour deux méthodes de pondération aux c.v. de la méthode PP, selon l'âge de la personne de référence



Rapports aux c.v. de la méthode PP



Les figures 2-4 sont la représentation graphique par treillis (Cleveland 1993) des valeurs c.v. et des quotients des moyennes pour calwts0 et calwts1. Le facteur calwts1 semble être la meilleure des diverses méthodes que nous avons examinées, au sens où elle améliore l'estimation de la rubrique «Toutes les dépenses», tout en offrant un rendement uniforme pour les différentes catégories de dépenses. Dans chaque partie des tracés, une ligne de référence verticale est tracée au point 1, c'est-à-dire le point d'égalité entre les résultats de calage et ceux obtenus pour la méthode PF. La gamme inférieure sur chaque tracé représente les quotients des moyennes des facteurs calwts0 et calwts1 sur les moyennes PF, et illustre le fait que, hormis quelques exceptions, les niveaux des moyennes obtenus par les deux méthodes de régression restreinte sont à peu près les mêmes que pour la méthode PF.

Les deux méthodes de calage choisies améliorent, dans l'ensemble, les c.v. par rapport à la méthode PF, c'est-à-d. que les quotients c.v. tendent à être inférieurs à 1 pour la plupart des domaines et des dépenses, le facteur calwts1 étant quelque peu meilleur à cet égard que le facteur calwts0. Pour les domaines d'âge de référence < 25 et 65+, par exemple, 12 des 15 catégories de dépenses présentent des quotients calwts1 inférieurs à 1. Pour les tailles des UC comprises entre 1 et 4, 9, 9 et 11. Il y a bien évidemment des exceptions. Pour la région du Sud, seulement six des 15 catégories de dépenses présentent, pour le facteur calwts1, des quotients c.v. égaux ou inférieurs à 1.

Les facteurs calwts2 et calwts3, qui utilisent le revenu familial avant impôt comme l'une des variables auxiliaires, présentent un rendement quelque peu erratique dans les différentes catégories, présentant parfois une amélioration importante par rapport à la méthode PF, mais, dans d'autres cas, des pertes majeures. Ce comportement est relié à la nature même de la variable revenu familial. Pour l'ensemble complet de données composé de 5,156 UC, le revenu avant impôt était positif pour 4,698 UC, nul pour 450 UC et négatif pour 8 UC. Les valeurs nulles représentent des quotients incomplets sur le revenu, tandis que les valeurs négatives sont associées à des familles pour lesquelles des pertes commerciales avaient été ajoutées à d'autres revenus. Dans un cas comme dans l'autre, ces UC entraînent l'utilité de cette variable pour la prévision des dépenses. Peut-être qu'en utilisant une autre mesure du revenu, combinée avec l'imputation des postes pour les revenus manquants, on pourrait améliorer la performance des facteurs calwts2 et calwts3 pour les estimations dans les différents domaines.

Compte tenu de tout ce qui précède, les facteurs regwts1, calwts1 et calwts4 sont des choix efficaces pour ce type d'application. Le facteur calwts1 a l'avantage, par rapport à regwts1, de ne pas offrir de valeur négative. Comme le calcul de calwts4 nécessite 23 variables auxiliaires, contre seulement 18 pour le facteur calwts1, ce dernier représente le choix le plus économique. À la suite de l'analyse que nous venons de présenter, nous avons effectué une étude similaire en utilisant

inférieur à 1, il y a alors amélioration par rapport à l'estimation PF, car, pour tous les facteurs de pondération, les estimations moyennes et les dépenses étaient très près de celles obtenues avec les estimations PF. Nous avons calculé les quotients des c.v. et les quotients des moyennes pour chacun des ensembles de facteurs de pondération décrits au tableau 1, pour chacune des dépenses choisies et pour chacun des domaines suivants :

- (1) Âge de la personne de référence: < 25, 25-34, 35-44, 45-54, 55-64, 65+;
- (2) Région: Nord-Est, Mid-Ouest, Sud, Ouest;
- (3) Taille de l'UC: 1, 2, 3, 4, 5+;
- (4) Composition des ménages: mari et femme seulement, mari et femme + enfants, mari et femme + autres, un parent + au moins un enfant < 18, personne seule et autres UC;
- (5) Type de ménage: propriétaire, locataire;
- (6) Race de la personne de référence: Noir, non-Noir.

Nous discuterons seulement des domaines (1) à (3) ici. En outre, les quotients pour toutes les UC, c'est-à-dire le total pour l'ensemble des domaines, ont été calculés pour chaque dépenses, et il sont présentés au tableau 2. Pour la catégorie «Toutes les dépenses», les facteurs regwts2, calwts2 et calwts3, avec des quotients de 0,79, 0,78 et 0,75 respectivement, présentent une réduction substantielle du c.v. par rapport à la méthode PF. Pour les dépenses ventilées, les facteurs regwts1 et calwts1 présentent une amélioration raisonnablement uniforme par rapport à la méthode PF, sans les pertes encourues par certains des autres facteurs pour des dépenses comme l'aménagement, les assurances personnelles et les pensions, ainsi que pour la sous-catégorie pensions et sécurité sociale.

Quotients des c.v. de la méthode PF sur les c.v. obtenus par les différentes méthodes de pondération.

Le quotient minimal dans chaque rangée est en gris

Dépense	regwts				calwts			
	0	1	2	0	1	2	3	4
Toutes les dépenses	0.98	0.90	0.79	0.98	0.90	0.78	0.75	0.87
Logement	0.93	0.85	0.75	0.93	0.85	0.74	0.72	0.84
Services publics	1.08	1.03	0.94	1.07	1.03	0.88	0.91	0.92
Aménagement	1.08	1.21	3.52	1.06	1.21	2.58	2.57	1.17
Gros électroménagers	1.08	1.06	1.04	1.06	1.08	1.09	1.00	1.03
Tous les véhicules	0.90	0.89	0.98	0.91	0.89	0.98	0.97	0.90
Automobiles et camions neufs	0.95	0.91	1.01	0.96	0.91	1.02	1.02	0.91
Automobiles et camions usagés	0.98	0.94	0.96	0.97	0.94	0.97	0.96	0.95
Essence, huile moteur	1.17	1.11	1.03	1.12	1.10	0.99	0.94	1.10
Soins de santé	1.05	0.97	0.86	1.07	0.97	0.85	0.87	0.94
Éducation	0.92	0.93	1.04	0.91	0.93	1.06	1.07	0.88
Contributions pédonnelles	1.01	1.02	1.28	1.01	1.02	1.30	1.29	1.03
Assurance personnelle, pensions	1.00	0.97	1.64	1.01	0.98	1.24	0.98	0.95
Assurances-vie et autres assurances	1.08	1.02	1.53	1.08	0.98	1.38	1.33	1.01
Pensions, sécurité sociale	1.00	0.99	1.75	1.01	0.99	1.34	1.06	0.97

(Les lignes de référence correspondent à  $L = 0.5$  et à  $U = 2$ )

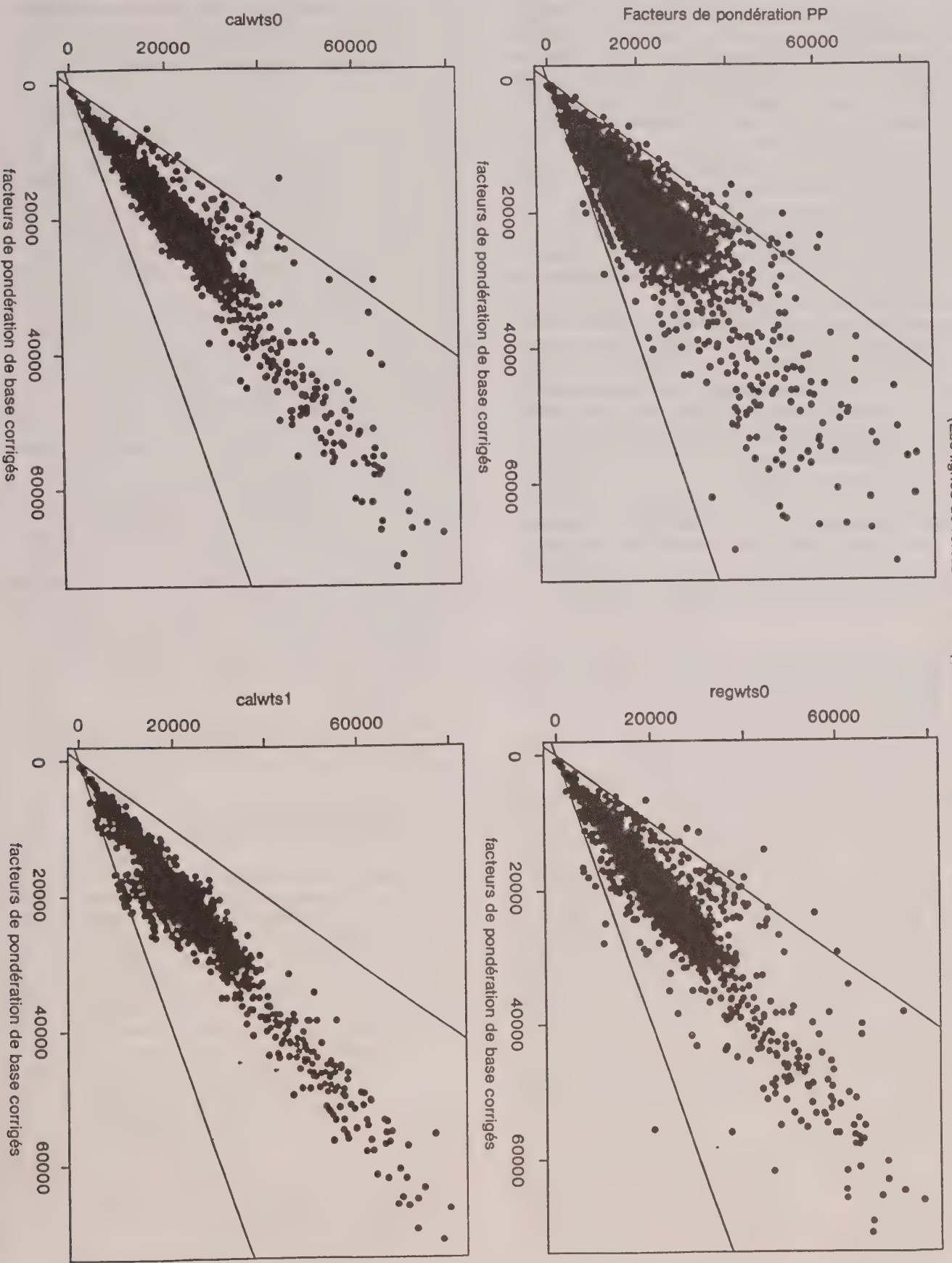


Figure 1. Graphiques de quatre ensembles de facteurs de pondération en fonction des facteurs de pondération de base corrigés







L'enquête CB recueille de l'information sur les habitudes de dépenses des consommateurs américains et le coût de la vie pour ceux-ci. L'enquête comporte deux parties, soit une entrevue trimestrielle et une enquête hebdomadaire par journal. L'enquête par entrevue recueille des données détaillées sur les types de dépenses dont il est probable que les répondants puissent se souvenir pour une période de trois mois ou plus (p. ex., propriété, automobiles, gros électroménagers) et qui représentent de 60 à 70% des dépenses totales du ménage. L'enquête par journal est réalisée à domicile par la famille enquêtée pendant deux périodes consécutives d'une semaine, et elle recueille des données sur toutes les dépenses de la famille pendant cette période. L'échantillon est choisi en deux étapes, avec les unités primaires d'échantillonnage géographiques à la première étape, et les ménages à la seconde.

Nous avons évalué les estimateurs décrits ci-dessus pour un certain nombre de dépenses tirées de l'enquête par entrevue. Nous avons utilisé les données recueillies pendant le deuxième trimestre de 1992, ce qui représentait  $n = 5156$  UC. L'unité primaire d'analyse, dans l'enquête CB, est l'unité de consommation, c'est-à-dire une famille économique au sein d'un ménage. Une unité de consommation (UC) consiste en personnes au sein du ménage qui partagent des dépenses. Ainsi, il peut y avoir plus d'une UC dans un ménage.

Nous avons étudié cinq ensembles différents de variables auxiliaires (les  $x_i$  dans la notation utilisée à la section 2). Nous avons choisi ces six variables en testant l'adéquation du modèle (2.1) avec les dépenses choisies pour différentes combinaisons des variables auxiliaires disponibles. Nous avons déterminé les combinaisons de variables auxiliaires pour lesquelles chaque coefficient de régression estimé était significatif dans une régression par les moindres carrés ordinaires, au niveau de 5%. Une étape simple mais importante qui a grandement amélioré l'ajustement avec le modèle a été d'inclure une coordonnée à l'origine. Nous avons également incorporé dans le choix des variables auxiliaires le fait que nous savions que l'enquête présentait un sous-dénombrement plus important des Noirs que des non-Noirs, et qu'il fallait en tenir compte par stratification à posteriori. Nous avons considéré cette méthode de choix des variables comme exploratoire et, par conséquent, nous avons étudié certains nombres de combinaisons afin de déterminer quel ensemble donnait les meilleurs estimateurs des dépenses moyennes. Nous avons inclus les 56 strates à posteriori basées sur l'âge, la race et le sexe, couramment utilisées dans l'enquête CB. (Ces 56 strates sont habituellement réduites, dans les activités réelles liées à l'enquête CB, en raison de la taille trop petite des échantillons dans certaines cellules.) Parmi les autres variables qui étaient statistiquement significatives dans diverses combinaisons, mentionnons la région (Nord-Est, Middle West, Sud et Ouest), l'urbanité (urbain/rural) par région, l'âge de la personne de référence de l'UC ( $< 25$ , 25-34, 35-44, 45-64, 65+), le type de ménage (propriétaire/locataire), le revenu de l'UC avant impôt et les 56 strates à posteriori groupées par sexe et certaines par catégories d'âge, afin de former 10 catégories d'âge/race. À partir de cette information, nous avons calculé les facteurs de pondération (2.8) en utilisant les valeurs données dans l'expression (2.9), c.-à-d. les valeurs regwts, et

pour calculer les totaux de contrôle des strates à posteriori. Ce désaccord entre l'unité d'analyse (le ménage) et l'unité de stratification à posteriori (la personne), lorsqu'une caractéristique du ménage présente un intérêt, a mené à l'élaboration de la méthode PP, utilisée dans les enquêtes CB (dépenses de consommation) et Current Population (enquêtes démographiques).

Dans la méthode PP décrite par Alexander (1987), on débute le processus de pondération en utilisant un facteur de pondération à base simple,  $a_i$ , que l'on corrige pour tenir compte des non-réponses. Le facteur de pondération corrigé est assigné à chaque personne dans le ménage, et les facteurs de pondération pour les personnes sont par la suite de nouveau corrigés afin que leur somme corresponde aux totaux de contrôle pour les personnes dans la population, par âge, race et sexe. Cette dernière correction peut donner à des personnes au sein d'un même ménage des pondérations différentes. On assigne ensuite au ménage le facteur de pondération de la personne désignée comme «personne principale» de ce ménage. Cette méthode comporte un certain élément d'arbitraire, et elle est difficile à analyser mathématiquement. Le but de cette recherche n'était pas de voir s'il était possible d'améliorer la méthode PP, mais plutôt d'utiliser la version actuelle de cette méthode comme base commune pour mesurer la performance d'autres estimateurs.

On peut formuler les estimateurs de régression (ordinaire et restreinte), de telle sorte que les totaux de contrôle pour les personnes dans la population soient satisfaits, toutes les personnes d'un même ménage ayant le même facteur de pondération, et sans qu'il soit nécessaire de choisir arbitrairement parmi les facteurs de pondération calculés pour les personnes afin d'assigner une pondération au ménage. On y parvient en définissant les variables auxiliaires au niveau du ménage. Par exemple, s'il y a trois strates d'âge à posteriori et que le ménage  $i$  compte 1, 0 et 2 personnes dans ces strates, le vecteur des données auxiliaires serait  $x_i = (1, 0, 2)'$ . On remarquera que cette formulation est différente de celle de Lemaître et Dufour (1987), qui ont défini les variables auxiliaires au niveau des personnes et ont assigné à chaque personne la moyenne des données pour le ménage, soit (1/3, 0, 2/3) dans l'exemple ci-dessus. Ces auteurs ont utilisé cette méthode par la «moyenne», parce qu'ils étaient intéressés à calculer des estimations à la fois pour les personnes (p. ex., le nombre de personnes employées) et pour les ménages (p. ex., les familles économiques). Nous n'avons besoin que d'un facteur de pondération pour le ménage, car nos variables cibles (c.-à-d.  $y_i$ ), comme les dépenses pour le logement ou les services publics, sont recueillies au niveau du ménage.

### 3. APPLICATION

Nous comparons les trois estimations (c.-à-d., régression, régression restreinte avec  $L = 0.5$  et  $U = 4$ , et personne principale) en les appliquant aux moyennes estimées et à leurs erreurs-types estimées pour un certain nombre de dépenses tirées de l'enquête CB, présentée par le Bureau of Labour Statistics.



On peut également exprimer l'estimateur de régression  $\hat{y}_R$  sous forme d'une somme pondérée des valeurs  $y_i$  de l'échantillon, ce qui est une caractéristique souhaitable pour les enquêtes. On peut voir aisément que l'expression (2.2) peut être réécrite sous la forme  $\hat{y}_R = \sum_{i \in s} w_i y_i$  où

$$w_i = a_i \left[ 1 + (X - \bar{x})' A^{-1} \frac{x_i}{\sigma_i^2} \right] \quad (2.5)$$

et où  $A = \sum_{i \in s} a_i x_i x_i' / \sigma_i^2$ . Les coefficients de pondération dépendent de l'échantillon par l'intermédiaire des  $x_i$  qui sont dans l'échantillon, mais cela est également vrai pour bon nombre d'estimateurs d'enquête, y compris l'estimateur de stratification a posteriori. Toutefois, ces facteurs de pondération ne dépendent pas de la variable  $y$  particulière étudiée, ce qui signifie que l'on peut utiliser un ensemble de facteurs de pondération  $w_i$  pour toutes les estimations.

On calcule de manière triviale une moyenne par unité, soit:  $\hat{y}_R = \bar{y}_R / N$  où  $\bar{y}_R = \sum_{i \in s} w_i y_i$ . Si nous calculons les totaux des valeurs auxiliaires  $x_i$ , alors

$$\sum_{i \in s} w_i x_i' = \sum_{i \in s} \left[ a_i x_i' + (X - \bar{x})' A^{-1} \frac{a_i x_i x_i'}{\sigma_i^2} \right] = X', \quad (2.6)$$

c'est-à-dire que nous reproduisons les totaux connus de la population. Ceci est également une caractéristique de l'estimateur de stratification a posteriori.

L'estimateur de  $\beta$  dans l'expression (2.4) ne tient compte d'aucune corrélation entre les erreurs dans le modèle (2.1). Dans les populations en grappes, il peut y avoir une corrélation entre les unités qui sont géographiquement près les unes des autres, p. ex., les UC d'un même quartier. L'utilisation d'une matrice des covariances  $V$  complète peut donner un résultat encore plus optimal (p. ex., voir Casady et Valliant 1993, et Rao 1994). Bien que l'utilisation d'une matrice des covariances  $V$  complète puisse réduire la variance de  $\beta$ , les éléments de  $V$  dépendront de la variable  $y$  particulière étudiée, et le calcul de  $V$  est habituellement pénible. Par conséquent, il est intéressant et commode de considérer le cas simple où  $V = \text{diag}(\sigma_i^2)$ , qui donne l'expression (2.2). On notera que lorsque la variance du plan de sondage  $\text{var}_p(\hat{y}_R)$  est estimée, il est nécessaire d'utiliser une méthode qui reflète de façon appropriée les grappes et les autres complexités du plan de sondage.

L'estimateur de régression a toutefois le désavantage que les facteurs de pondération peuvent avoir des valeurs beaucoup trop grandes ou faibles, voire négatives. Les estimateurs de calage ressemblent à Särndal (1992), que nous présentons maintenant, ajoutent des contraintes afin de limiter la taille des facteurs de pondération. Les estimateurs de calage sont formés par minimisation d'une distance donnée,  $F$ , entre un facteur de pondération initial et le facteur de pondération final, compte tenu de certaines contraintes. Ces contraintes peuvent comporter les variables auxiliaires disponibles, et ces dernières se trouvent donc incorporées dans l'estimateur. L'estimateur de régression présente ci-dessus est un cas spécial de l'estimateur de calage dans lequel  $F$  est définie

comme étant la fonction de distance par les moindres carrés

$$F(w_i a_i) = \frac{a_i c_i}{w_i} \left( \frac{2}{a_i} - 1 \right)^2$$

pour  $i = 1, \dots, n$ , les valeurs  $c_i$  étant un facteur de pondération positif connu (p. ex., si  $c_i = \sigma_i^2$  ou  $c_i = 1$ ) associé à l'unité  $i$ , et  $w_i$  le facteur de pondération final. La distance totale de de l'échantillon  $\sum_{i \in s} F(w_i a_i)$  est minimisée, selon les contraintes

$$\sum_{i \in s} w_i x_i' = X. \quad (2.7)$$

Sous cette forme, on peut écrire les facteurs de pondération de l'estimateur de régression pour la population totale de  $y$  indiquée dans (2.5), comme suit :

$$w_i = a_i g(c_i^{-1} \lambda' x_i) \quad \text{pour } i = 1, \dots, n, \text{ où} \quad (2.8)$$

$$g(u) = 1 + u,$$

(2.9)

pour  $n \in \mathcal{H}$  et  $\lambda$  est un multiplicateur de Lagrange calculé dans le processus de minimisation. La forme particulière de  $w_i$  avec

$c_i = \sigma_i^2$  pour l'estimateur de régression a été donnée dans (2.5). Afin d'éliminer les extrêmes, on peut définir les facteurs de pondération en restreignant  $g$  de telle sorte que:

$$g(u) = \begin{cases} L & \text{si } u < L - 1 \\ 1 + u & \text{si } L - 1 \leq u \leq U - 1 \\ U & \text{si } u > U - 1 \end{cases} \quad (2.10)$$

Avec cette définition de  $g$ , les facteurs de pondération  $w_i$  satisfont à l'expression

$$L < w_i a_i < U \quad (2.11)$$

Dans la plupart des enquêtes-ménages, la stratification a posteriori sert avant tout à corriger le sous-dénombrement de la population cible à cause de la base de sondage et de l'échantillon. Aux États-Unis, il y a peu de dénominations fiables des ménages dans la population qui sont utilisables pour la stratification a posteriori. Par conséquent, on utilise habituellement les chiffres de personnes dans la population

l'estimation par régression est que bon nombre des techniques standard d'enquêtes, y compris l'estimateur de stratification a posteriori mentionné ci-dessus, sont des cas spéciaux des estimateurs de régression. En outre, les estimateurs de régression incorporent avec plous de souplesse les données auxiliaires que les autres méthodes acceptent facilement dans une enquête-ménage, cette méthode accepte facilement les données auxiliaires au niveau des personnes et au niveau des ménages, qu'elles soient de nature qualitative ou quantitative. Parmi les travaux portant sur l'estimation par régression et la stratification a posteriori, mentionnons ceux de Bethlehem et Keller (1987), Casady et Valliant (1993), Deville et Särndal (1992), Deville, Särndal et Sautory (1993) et Zieschang (1990).

Dans la présente étude, nous comparons l'estimateur de régression avec l'estimateur PP actuellement utilisé au Bureau of Labour of Statistics (BLS). On peut écrire chaque estimateur sous forme d'une somme pondérée des valeurs des variables de réponse dans l'échantillon. Chaque facteur de pondération est ensuite interprété de façon classique comme étant le nombre de personnes dans la population qui aurait la valeur correspondante de la variable de réponse. Cette interprétation demande que chaque facteur de pondération soit égal ou supérieur à un. L'estimateur de régression par les moindres carrés ordinaires présente un désavantage, car il peut donner des facteurs de pondération non positifs. Un certain nombre d'auteurs ont suggéré des méthodes pour surmonter ce problème. La méthode la plus facile est probablement celle qui a été présentée par Deville et Särndal (1992), qui consiste à éliminer tout facteur de pondération négatif, et à éliminer également les facteurs de pondération extrêmes. Les estimateurs de régression restreinte, obtenus au moyen de ces nouveaux facteurs de pondération, sont également comparés à l'estimateur de régression original et à l'estimateur PP.

Nous décrivons les trois différents estimateurs à la section 2. La section 3 présente une application de ces procédures, dans le cadre de l'enquête sur les dépenses de consommation faite par le BLS – le même contexte que celui qui est décrit par Zieschang (1990). Nous comparons les coefficients de variation pour un certain nombre de variables cibles de l'enquête, pour la population entière et pour un certain nombre de domaines. La section 4 est un résumé de nos conclusions.

## 2. ESTIMATEURS DE RÉGRESSION, DE CALAGE ET DE PERSONNE PRINCIPALE

Tout d'abord, présentons brièvement l'estimateur de régression. Un échantillon  $s$  de taille  $n$  est choisi dans une population finie  $U$  de taille  $N$ . Supposons que la probabilité de choisir la  $i$ -ième unité soit  $\pi_i$ . L'échantillon peut être à deux degrés, et l'unité peut être soit l'unité primaire d'échantillonnage, soit l'unité secondaire d'échantillonnage. Nous n'avons nul besoin ici de compliquer la notation en introduisant des indices explicites pour chaque degré de l'échantillonnage. Dénotons la variable qui nous intéresse par  $y$  et supposons que sa valeur dans la  $i$ -ième unité,

$y_i$ , soit observée pour chaque  $i \in s$ . Supposons qu'il existe  $K$  variables auxiliaires  $x_1, x_2, \dots, x_K$  dont nous connaissons les valeurs pour chaque  $i \in s$ . Définissons  $x_i = (x_{i1}, x_{i2}, \dots, x_{iK})'$ , pour chaque  $i \in U$ , où  $x_{ik}$  représente la valeur de la variable  $x_k$  dans l'unité  $i$ . Représentons par  $X = (X_1, \dots, X_K)'$  le vecteur de dimension  $K$  des totaux connus de la population pour les variables  $x_1, x_2, \dots, x_K$ . L'estimateur de régression est donc motivé par le modèle de travail  $\xi$  suivant:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_K x_{iK} + \varepsilon_i \quad (2.1)$$

pour  $i = 1, \dots, N$ . Ici,  $\beta_1, \dots, \beta_K$  sont des paramètres inconnus du modèle. Les  $\varepsilon_i$  sont des erreurs aléatoires, pour lesquelles  $E_\xi(\varepsilon_i) = 0$  et  $\text{var}_\xi(\varepsilon_i) = \sigma_i^2$  pour  $i = 1, \dots, N$ . Nous utilisons l'expression «modèle de travail» afin de souligner le fait que le modèle est probablement erroné jusqu'à un certain point. Dans l'enquête sur les dépenses de consommation (CB), l'unité d'analyse, indexée par  $i$ , est une unité de consommation (UC), qui est similaire à un ménage et qui est définie plus en détail à la section 3. La valeur  $y_i$  peut être les dépenses alimentaires totales par les UC et les  $x_{ik}$  peuvent être diverses caractéristiques des UC, par exemple le nombre de personnes d'âges différents ou le revenu de l'UC. La variance des dépenses peut dépendre de la taille de l'UC, de telle sorte qu'il serait raisonnable que l'expression  $\sigma_i^2$  soit proportionnelle au nombre de personnes dans l'UC. Nous incluons une coordonnée à l'origine dans certains de nos modèles, en mettant la première variable auxiliaire  $x_1$  égale à 1. On définit comme suit un estimateur de régression linéaire pour la population totale  $y$ :

$$y_R = y^\pi + (X - X^\pi)' \beta \quad (2.2)$$

où  $y^\pi$  dénote le  $\pi$ -estimateur (ou estimateur de Horvitz-Thompson) de la population totale de  $y$ , c.-à-d.

$$y^\pi = \sum_{i \in s} a_i y_i \quad (2.3)$$

où  $a_i = 1/\pi_i$ . En outre,  $X^\pi = (x_1^\pi, \dots, x_K^\pi)'$  est le vecteur des  $\pi$ -estimateurs des totaux de population pour les variables  $x_1, x_2, \dots, x_K$  et

$$\beta = (\beta_1, \dots, \beta_K)' = \left[ \sum_{i \in s} a_i x_i x_i' / \sum_{i \in s} a_i^2 \right]^{-1} \sum_{i \in s} a_i x_i y_i / \sum_{i \in s} a_i^2 \quad (2.4)$$

Nous supposons que l'expression  $\sum_{i \in s} a_i x_i x_i' / \sigma_i^2$  est non singulière. Même si le modèle (2.1) est inflexible jusqu'à un certain point,  $y_R/N$  est un estimateur cohérent avec le plan de sondage pour la moyenne  $\bar{y}$  de la population, peu importe que le modèle supposé soit vrai ou faux. Cela est évident d'après l'expression (2.2). Si  $y^\pi/N$  et  $X^\pi/N$  sont des estimateurs convergent avec le plan de sondage de  $\bar{y}$  et de  $\bar{X}$ , qui est le vecteur des moyennes de la population des variables auxiliaires, alors le deuxième terme dans  $y_R/N$  converge vers zéro, tandis que le premier converge vers  $\bar{y}$ . Pour plus de détails, voir Särndal, Swensson et Wretling (1992).



# Application de l'estimation par régression restreinte dans une enquête-ménage

BODHINI R. JAYASURIYA et RICHARD VALLIANT<sup>1</sup>

## RÉSUMÉ

Dans cet article, les auteurs comparent empiriquement trois méthodes d'estimation – par régression, par régression restreinte et au moyen de la méthode dite de la personne principale – utilisées dans une enquête-ménage sur les dépenses de consommation. Les trois méthodes sont appliquées à la stratification a posteriori, qui est importante dans de nombreuses enquêtes-ménages afin de corriger le sous-dénombrement de la population cible. Dans les recensements externes, on dispose habituellement de chiffres de population pour des strates a posteriori pour les personnes, mais non pour les ménages. Si on a besoin d'estimations par ménage, on doit assigner un facteur de pondération unique à chaque ménage, tout en utilisant le nombre de personnes pour la stratification a posteriori. On y parvient facilement en employant des estimateurs de régression pour les totaux ou les moyennes, et en utilisant le nombre de personnes dans les données auxiliaires de chaque ménage. L'estimation par régression restreinte permet de mieux calculer les facteurs de pondération, car on contrôle les valeurs extrêmes et l'on peut obtenir des estimateurs présentant une variance moindre que les estimateurs de Horvitz-Thompson, tout en respectant les totaux de contrôle de la population. Les méthodes de régression permettent également d'utiliser des contrôles pour les chiffres au niveau des personnes et des ménages et pour les données auxiliaires quantitatives. Avec la méthode dite de la personne principale, les personnes sont classées dans les strates a posteriori, et les facteurs de pondération pour les personnes font l'objet d'un rajustement par quotient afin d'obtenir des totaux de contrôle de la population. De la sorte, chaque personne dans un ménage peut se voir attribuer un facteur de pondération différent. Le facteur de pondération associé à la «personne principale» est alors choisi comme facteur de pondération pour le ménage. Nous comparerons les moyennes calculées à partir des trois méthodes, ainsi que leurs erreurs-types estimées, pour un certain nombre de dépenses tirées de l'enquête sur les dépenses de consommation parrainée par le Bureau of Labor Statistics.

MOTS CLÉS: Calage; méthode de la personne principale; variance de répétition; régression restreinte.

## 1. INTRODUCTION

Dans les grandes enquêtes-ménages, le sous-dénombrement de la population cible constitue un problème de signal, qui est souvent attribuable à des taux de réponse différents parmi les sous-groupes de la population et à des lacunes dans la base de sondage. La stratification a posteriori est une méthode utilisée à l'étape de l'estimation afin de réduire les erreurs quadratiques moyennes basées sur l'information qui influe sur les variables de réponse. L'estimateur est construit de telle façon que le nombre total estimé de personnes s'inscrive dans chaque strate a posteriori soit égal au chiffre de population véritable. Les chiffres de population par strate a posteriori proviennent habituellement d'un recensement externe portant sur le nombre de personnes, mais pas toujours sur le nombre de ménages. Si on a besoin d'estimations par ménage, il faut assigner un facteur de pondération unique à chaque ménage, tout en utilisant les chiffres de personnes pour la stratification a posteriori. On y parvient en estimant par régression les valeurs totales ou moyennes portant sur les chiffres de personnes dans les données auxiliaires de chaque ménage. L'estimation par régression restreinte contrôle les facteurs de pondération extrêmes, et peut fournir des estimateurs dont la variance est moindre que celles de l'estimateur de Horvitz-Thompson, tout en permettant de respecter les totaux de contrôle de la population. Une autre méthode utilisée dans certaines enquêtes est la méthode dite de la personne

*principale* (PP) (Alexander 1987), dans laquelle le facteur de pondération du ménage est basé sur la personne désignée comme «personne principale» dans chaque ménage. Les personnes sont classées dans des strates a posteriori, et les facteurs de pondération des personnes font l'objet d'un rajustement des quotients afin d'obtenir des totaux de contrôle de la population, ce qui permet d'attribuer à chaque personne dans un ménage un facteur de pondération différent. Le facteur de pondération associé à la personne principale est ensuite assigné au ménage. Cette méthode spéciale est difficile à analyser théoriquement. Bien qu'on puisse facilement corriger pour tenir compte du sous-dénombrement de la population, les estimateurs de régression qui font l'objet du présent article fournissent automatiquement un facteur de pondération pour les ménages qui n'est basé sur aucun de leurs membres particuliers. Lemaitre et Dufour (1987) ont traité de l'utilisation, par Statistique Canada, de l'estimateur de régression dans un tel contexte. On constate une utilisation croissante des estimateurs de régression dans les enquêtes, tant dans les travaux théoriques que dans les enquêtes réelles. Statistique Canada a incorporé l'estimateur de régression général dans son logiciel Système généralisé d'estimation (GES), qui est maintenant utilisé dans bon nombre de ses enquêtes (Estevao, Hidiroglou et Særdal 1995). Fuller, Louglin et Baker (1993) ont traité d'une application de cette méthode dans l'enquête américaine sur la consommation alimentaire (USDA). Un des attraits de

<sup>1</sup> Bodhini R. Jayasuriya et Richard Valliant, U.S. Bureau of Labor Statistics, 2 Massachusetts Avenue, N.E., Room 4915, Washington, DC 20212, U.S.A.

- SINGH, M.P., DREW, J.D., GAMBINO, J.G., et MAYDA, F. (1990). *Méthodologie de l'enquête sur la population active du Canada: 1984-1990*. N° 71-526 au catalogue, Statistique Canada.
- STUKEL, D.M., et BOYER, R. (1992). Calibration estimation: An application to the Canadian Labour Force Survey. Documents de travail de la Direction de la méthodologie, SSMD-92-009E, Statistique Canada.
- WESVARPC (1995). Westat Inc., Rockville, Maryland. travail de la Direction de la méthodologie, SSMD-92-009E, Statistique Canada.
- WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.
- YUNG, W., et RAO, J.N.K. (1996). Linéarisation des estimateurs de variance jackknife dans un échantillonnage stratifié à degrés multiples. *Techniques d'enquête*, 22, 23-31.
- HUANG, E.T., et FULLER, W.A. (1978). Nonnegative regression estimation for sample survey data. *Proceedings of the Social Statistics Section, American Statistical Association*, 300-305.
- KOVAČEVIĆ, M.S., YUNG, W., et PANDHER, G.S. (1995). Estimating the sampling variances of measures of income inequality and polarization – An empirical study. Documents de travail de la Direction de la méthodologie, HSMMD-95-007E, Statistique Canada.
- SÄRNDAL, C.-E. (1982). Implications of survey design for generalized regression estimation of linear functions. *Journal of Statistical Planning and Inference*, 7, 155-170.
- SÄRNDAL, C.-E., SWENSSON, B., et WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- SINGH, A.C., et MOHL, C.A. (1996). Comprendre les estimateurs de calage dans les enquêtes par échantillonnage. *Techniques d'enquête*, 22, 107-116.



inversement corrélés au nombre d'UPB tirées. On a relevé e nombre d'UPB en faisant en sorte qu'il y ait autopondération du plan d'échantillonnage; cette approche semble avoir eu les plus grands effets sur la réduction des CV. Le second degré du plan d'échantillonnage n'a pas été modifié. La deuxième simulation a eu pour avantage secondaire de doubler approximativement la taille de l'échantillon, ce qui nous a permis de surmonter les difficultés de convergence mentionnées au paragraphe précédent.

Les résultats de la deuxième simulation apparaissent au tableau 2. La dernière colonne indique la diminution du nombre d'échantillons écartés en raison d'un problème de convergence. Les cinquante et sixième colonnes montrent que les CV ont sensiblement diminué pour se situer entre 22,70% et 24,2%, les estimateurs jackknife se caractérisant toujours par une valeur légèrement plus élevée. Comme c'était le cas auparavant, le biais relatif en pour cent de l'estimateur ponctuel reste négligeable, puisqu'il se maintient toujours nettement sous un pour cent. Lors de la simulation antérieure, le biais relatif en pour cent de l'estimateur de Taylor s'établissait constamment autour de -6%; lors de la nouvelle simulation, il fluctue toujours autour de -3%, ce qui signifie une fois de plus que la variance réelle a été sous-estimée. Encore une fois, on note très peu de variation dans le biais résultant des équations (3.1) et (3.2) pour la fonction de distance MCG. Le biais relatif en pour cent de l'estimateur jackknife (toujours d'environ -1,5%) est constamment inférieur en valeur absolue à celui de l'estimateur de Taylor (valeur absolue). Il existe cependant un cas (IQ restrictive ( $L = 0,8$ ,  $U = 1,3$ )) où l'estimateur jackknife soulève un problème de convergence; on n'a pas tenu compte de ces résultats, identifiés par un «\*». Fait surprenant, on ne remarque pratiquement aucun changement de biais pour les fonctions de distance restrictives avec les estimateurs de Taylor et jackknife, car les limites deviennent successivement plus étroites. En réalité, le biais relatif en pour cent des diverses fonctions de distance, tant pour l'estimateur de Taylor que pour l'estimateur jackknife, ne paraît guère varier. Signalement qu'à la deuxième simulation, l'erreur de Monte Carlo variait entre 0,37% et 2,13%.

## 5. CONCLUSIONS

Dans ce document, nous avons étudié le comportement des estimateurs ponctuels et des estimateurs de la variance de Taylor et jackknife correspondants pour plusieurs fonctions de distance auxquelles la théorie de calage donne accès. On a particulièrement insisté sur les fonctions de distance qui restreignent la fourchette de valeurs des coefficients  $g$ , donc éliminent un poids final éventuellement négatif ou positif mais très élevé. Tous les estimateurs ponctuels examinés se caractérisaient par un biais négligeable. Les estimateurs jackknife et de Taylor révèlent une faible sous-estimation de la variance réelle, bien que le biais de l'estimateur jackknife soit toujours inférieur à celui de l'estimateur de Taylor (valeur absolue). Le résultat le plus étonnant demeure que, pour les estimateurs de Taylor et jackknife, le biais ne varie pratiquement pas, aussi bien quand on limite à

l'extrême le coefficient  $g$  que dans le cas de limites moins étroites. Il conviendrait néanmoins de se montrer prudent lorsqu'on utilise des limites extrêmes, étant donné les difficultés de convergence qui peuvent surgir, surtout lorsqu'on recourt à la méthode jackknife pour estimer la variance, les estimateurs ponctuels devant être recalculés constamment. Si l'usage des fonctions de distance restrictives a pour but principal d'éliminer un poids éventuellement négatif ou positif très élevé, il suffirait d'imposer des limites modestes aux coefficients  $g$ .

En guise de remarque finale, soulignons qu'il est intéressant de constater que l'estimateur de jackknife a réclarné environ 97% du temps de calcul alors que la méthode de Taylor linéarisée n'en a exigé que les 3% restants. Pareille différence, qu'on pourrait qualifier d'extrême, dans le temps nécessaire au calcul pourrait avantager la méthode de Taylor si on a besoin de mesurer précisément un grand nombre de dimensions. Face aux progrès récents réalisés au niveau de l'efficacité des calculs relatifs à l'estimateur de la variance jackknife (programme WESVARPC (1995), par exemple), il se pourrait qu'on puisse compenser ce déséquilibre. Malgré cela, rappelons qu'à l'heure actuelle, le programme WESVARPC n'a amélioré l'efficacité de calcul que des plans d'échantillonnage à deux UPB par strate dont les estimateurs de stratification a posteriori ne présentent qu'une dimension. En conclusion, nos travaux ne démontrent pas de façon concluante la supériorité d'un estimateur de la variance quelconque et indiquent que les deux estimateurs fonctionnent raisonnablement bien pour toutes les fonctions de distance. Il revient donc à l'utilisateur de déterminer la combinaison de variance et de fonction de distance qui fonctionne le mieux, compte tenu des exigences du système.

## REMERCIEMENTS

Les auteurs tiennent à remercier Chris Mohl qui leur a fourni une partie des codes informatiques utilisés lors de la simulation. Ils aimeraient également remercier le rédacteur en chef associé et deux examinateurs pour leurs remarques utiles sur la première version du document.

## BIBLIOGRAPHIE

- DEMING, W.E., et STEPHAN, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11, 427-444.
- DEVILLE, J.-C., et SÄRNDAAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- DEVILLE, J.-C., SÄRNDAAL, C.-E., et SAUTORY, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.
- HIDROGLOU, M.A., FULLER, W.A., et HICKMAN, R.D. (1980). SUPERCAR, Department of Statistics, Iowa State University, Ames, Iowa.

Tableau 2  
Biases relatif en pour cent des estimateurs ponctuels, et biais relatif en pour cent et CV en pour cent des estimateurs de la variance de Taylor et jackknife (échantillon d'environ 2000)

Fonction de distance	Biases relatif en %	Biases relatif en %	Biases relatif en %	CV en %	CV en %	CV en %	Nombre d'échantillons (sur 4000)
	estimeur ponctuel	variance Taylor	variance jackknife	variance Taylor	variance jackknife	variance jackknife	
MCG (régression)	.02	-2.71 (ég 3.1)	-1.43	23.03 (ég 3.1)	22.84 (ég 3.2)	23.29	0
MCG restrictive	( $L = 0, U = 4$ ) ( $L = .4, U = 2$ ) ( $L = .68, U = 1.6$ ) ( $L = .8, U = 1.3$ )	-2.61 -.261 .02 .02	-1.43 -1.43 -1.44 -1.56	22.84 22.84 22.84 22.70	22.84 22.84 22.84 22.70	23.29 23.29 23.29 23.15	0 0 0 118
Itération du quotient	.25	-2.75	-1.15	22.84	22.84	23.43	0
IQ restrictive	( $L = 0, U = 4$ ) ( $L = .4, U = 2$ ) ( $L = .68, U = 1.6$ ) ( $L = .8, U = 1.3$ )	.17 .16 -.27 .27	-1.36 -1.42 -0.49 *	22.84 22.84 22.83 22.70	22.84 22.84 22.83 22.70	23.30 23.29 24.20 *	0 0 0 118
Huang-Fuller modifiée	( $L = 0, U = 4$ ) ( $L = .4, U = 2$ ) ( $L = .68, U = 1.6$ ) ( $L = .8, U = 1.3$ )	.02 -.261 -.261 .02	-1.43 -1.43 -1.44 -1.36	22.84 22.84 22.84 22.73	22.84 22.84 22.84 22.73	23.29 23.29 23.29 23.18	0 0 0 116
Rétrécissement-minimisation	( $L = 0, U = 4$ ) ( $L = .4, U = 2$ ) ( $L = .68, U = 1.6$ ) ( $L = .8, U = 1.3$ )	.02 -.261 -.261 .02	-1.43 -1.43 -1.44 -1.24	22.84 22.84 22.84 22.73	22.84 22.84 22.84 22.73	23.29 23.29 23.29 23.63	0 0 0 118

asymptotique des estimateurs de calage par rapport à l'esti-

mateur de régression.

La troisième colonne indique le biais relatif en pour cent

de l'estimateur de la variance de Taylor. Dans ce cas, on sous-

estime toujours la variance réelle, mais jamais de plus de

6.2%. En ce qui concerne l'estimateur de régression, l'in-

clusion du coefficient  $g$  à la formule de la variance (équation

(3.1) ou équation (3.2)) ne paraît guère faire de différence; le

biais ne s'améliore que légèrement lorsqu'on ajoute le

coefficient  $g$  (-5.82% contre -6.01%). L'estimateur de la

variance jackknife (quatrième colonne), par contre, donne

constamment de meilleurs résultats que l'estimateur de la

variance de Taylor. Ainsi, il sous-estime presque toujours la

variance réelle, mais de moins de 2% dans tous les cas.

Toutes les fonctions de distance aboutissent à une solution,

mais la fonction MCG exige un algorithme d'itération. À

cause de cela, quelques-uns des 4,000 échantillons posent des

problèmes de convergence, surtout aux limites extrêmes du

coefficient  $g$ . On a donc rejeté les échantillons pour lesquels

l'algorithme ne débouchait pas sur la convergence. Par

conséquent, ces échantillons n'ont pas contribué aux diverses

mesures de Monte Carlo. La dernière colonne du tableau 1

indique le nombre d'échantillons écartés. Aux limites

extrêmes ( $L = 0.68, U = 1.6$  et  $L = 0.8, U = 1.3$ ), on a dû en

rejeter tant (de 231 à 234 pour  $L = 0.68, U = 1.6$  et de 1,562

à 1,602 pour  $L = 0.8, U = 1.3$ ) que les résultats ne nous ont

pas paru fiables. Ils ne sont donc pas signalés. Ces limites

plus étroites ne sont néanmoins pas dépourvues d'intérêt.

C'est la raison pour laquelle on a repris la simulation en

doublant à peu près la taille de l'échantillon (hausse d'environ

1,000 à 2,000). Remarquons que Deville et Särndal (1992)

montrent que pour toutes les fonctions de distance, la

convergence est atteinte avec probabilité égale à un si la taille

de l'échantillon est suffisante.

Les colonnes cinq et six du tableau 1 indiquent les CV en

pour cent des estimateurs de la variance de Taylor et

jackknife. Les coefficients de variation de toutes les fonctions

de distance se ressemblent, puisque leurs valeurs fluctuent de

59.45% à 64.21%. Toutefois, ceux de la méthode jackknife

sont toujours un peu au-dessus de ceux de la méthode de

Taylor. Sans nier leur importance, on a déjà observé des

coefficients de variation d'un tel ordre de grandeur dans

d'autres simulations se rapportant à la variance. Lire par

exemple Kováčević, Yung et Pandher (1995). Nous voulions

cependant savoir si les principaux résultats concernant le biais

des estimateurs de la variance tiendraient toujours advenant

une réduction des coefficients de variation. À la suggestion

d'un examinateur, nous avons donc effectué une autre

simulation en portant le nombre d'UPÉ de 47 à 83, sachant

que les CV des estimateurs de la variance sont à peu près



**Tableau 1**  
Biais relatif en pour cent des estimateurs ponctuels, et biais relatif en pour cent et coefficients de variation en pour cent des estimateurs de la variance de Taylor et jackknife (échantillon d'environ 1000)

Fonction de distance	Biais relatif en % estimateur ponctuel	Biais relatif en % variance Taylor	Biais relatif en % variance Taylor	Biais relatif en % variance jackknife	CV en % variance Taylor	CV en % variance jackknife	Nombre d'échantillons écarts (sur 4000)
MCG (régression)	.11	-6.01 (ég 3.1)	-1.73	60.79 (ég 3.1)	59.60 (ég 3.2)	62.86	0
MCG restrictive	(L = 0, U = 4) .11	-5.82	-1.73	59.60	59.93	62.86	0
	(L = .4, U = 2) .10	-5.36	-1.27	59.60	59.94	63.21	32
Itération du quotient	.52	-6.20	0.84	59.45	59.45	63.35	0
IQ restrictive	(L = 0, U = 4) .50	-6.09	-0.31	59.48	59.81	63.47	0
	(L = .4, U = 2) .46	-5.69	-0.39	59.81	59.81	64.21	32
Huang-Fuller	(L = 0, U = 4) .11	-5.82	-1.73	59.60	59.94	62.86	0
	(L = .4, U = 2) .10	-5.36	-1.20	59.94	59.94	63.27	32
Rétrécissement-minimisation	(L = 0, U = 4) .11	-5.82	-1.73	59.60	59.94	62.86	0
	(L = .4, U = 2) .10	-5.36	-1.27	59.94	59.94	63.25	32

ou

$$(4.2) \quad \frac{(E_M(\hat{V}(\hat{Y}_w)) - V_{true})}{V_{true}} * 100$$

ou

$$E_M(\hat{V}(\hat{Y}_w)) = \frac{1}{R} \sum_{r=1}^R \hat{V}_r(\hat{Y}_w)$$

et

$$V_{true} = \frac{1}{R} \sum_{r=1}^R (\hat{Y}_w - E_M(\hat{Y}_w))^2$$

et  $\hat{V}_r(\hat{Y}_w)$  correspond à la valeur de  $\hat{V}(\hat{Y}_w)$  (Taylor ou jackknife) de l'échantillon  $r$ .

(C) Le coefficient de variation en pour cent de l'estimateur de la variance de Taylor/jackknife (par rapport à la variance réelle) est estimé par

$$(4.3) \quad \sqrt{\frac{\frac{1}{R} \sum_{r=1}^R (\hat{V}_r(\hat{Y}_w) - V_{true})^2}{V_{true}}} * 100$$

c'est-à-dire l'erreur quadratique moyenne de l'estimateur de la variance divisée par la variance réelle en pour cent. Bien que la plupart des études portent sur le biais des estimateurs de la variance, il vaut la peine d'examiner le coefficient de variation des estimateurs de la variance pour savoir dans quelle mesure les estimations de la variance fluctuent elles-mêmes.

## 4.2 Résultats de l'étude

Remarquons qu'il aurait pu s'avérer préférable de comparer les valeurs à l'«erreur quadratique moyenne réelle» plutôt qu'à la «variance réelle» aux équations (4.2) et (4.3). Toutefois, dans notre simulation, le biais relatif était si faible que l'écart entre les deux comparaisons est à toute fin pratique négligeable.

Enfin, pour déterminer si le nombre d'échantillons retenu était approprié, on a calculé l'erreur de Monte Carlo en prenant le coefficient de variation en pour cent de  $E_M(\hat{V}(\hat{Y}_w))$  donné par:

$$(4.4) \quad \sqrt{\frac{\frac{1}{R} \sum_{r=1}^R [\hat{V}_r(\hat{Y}_w) - E_M(\hat{V}(\hat{Y}_w))]^2}{E_M(\hat{V}(\hat{Y}_w))}} * 100.$$

Avec  $R = 4,000$ , l'erreur de Monte Carlo était constamment faible (entre 0.99% et 3.60%) pour les méthodes du jackknife et de Taylor, signe que les résultats sont stables.

Le tableau 1 indique le biais relatif en pour cent de l'estimateur ponctuel (équation (4.1)) et des estimateurs de la variance de Taylor et jackknife (équation (4.2)), ainsi que les coefficients de variation en pour cent des mêmes estimateurs (équation (4.3)). Le biais relatif en pour cent des estimations ponctuelles (colonne deux), soit beaucoup moins que 1% dans tous les cas. Le fait que les estimations ponctuelles présentent un biais similaire est raisonnable, étant donné l'équivalence

survenue en 1991, dans Singh, Drew, Gambino et Mayda (1990). Bricèvement, précisons que les provinces sont stratiées en «régions économiques», soit de vastes zones de structure économique analogue. Terre-Neuve compte quatre régions économiques. Ces dernières sont subdivisées en «unités autoréprésentatives» (UAR) et «unités non autoréprésentatives» (UNAR), elles-mêmes scindées en strates de niveau inférieur. Les UAR correspondent aux agglomérations de plus de 15,000 habitants comme St. John's et Cornerbrook, à Terre-Neuve. Au niveau le plus détaillé, Terre-Neuve compte 45 strates, comprenant chacune moins de six unités primaires d'échantillonnage (UPF), un nombre insuffisant pour l'échantillonnage dans le cadre de notre simulation. Les 45 strates ont donc été regroupées en 18, renfermant chacune de six à 18 UPF. Le regroupement n'a pas modifié la région économique, ni les régions métropolitaines de recensement (RMR) de St. John's et de Cornerbrook.

Pour la simulation de Monte Carlo, on a tiré  $R = 4,000$  échantillons d'environ 1,000 unités dans la «population» de Terre-Neuve (égale à 9,152) en vertu d'un plan d'échantillonnage à deux degrés. Des strates groupées venant des UNAR, on a sélectionné deux UPF par échantillonnage avec remise avec probabilité proportionnelle à la taille au premier degré. La taille de l'échantillon correspondait au nombre de logements de l'UPF. Au deuxième degré, on a retenu un logement sur cinq dans les UPF sélectionnées, par échantillonnage aléatoire simple à tirage sans remise. Pour les strates groupées appartenant à des UAR, on a choisi trois UPF par échantillonnage avec remise à la première étape avec probabilité proportionnelle à la taille. Au second degré, tous les logements des UPF échantillonnées ont été retenus, si bien que cette partie du plan d'échantillonnage se bornait à un échantillonnage de grappes complètes. Cette solution s'est avérée nécessaire faute d'un nombre suffisant de logements par UPF pour procéder à un sous-échantillonnage dans les UAR. Le choix de deux UPF dans les strates UNAR plutôt que trois, comme dans les strates UAR, est attribuable au fait qu'en général, les premières comptent moins d'UPF que les secondes. Dans l'ensemble, 47 UPF ont été échantillonnées. Dans l'un ou l'autre cas (UNAR ou UAR), tous les habitants des logements se sont tous retrouvés dans l'échantillon. Bien que ce plan soit un amalgame des plans à un et à deux degrés, nous le considérerons comme un plan d'échantillonnage à deux degrés pour plus de commodité.

- (1) la fonction de distance généralisée des moindres carrés (MCG) (équation (2.4)),
- (2) la fonction de distance par itération du quotient (IQ) (équation (2.7)),
- (3) la fonction de distance MCG restrictive (MCGR) (équation (2.8)),

Nous nous sommes intéressés à  $X$ , le nombre total de chômeurs. Cette valeur a été obtenue par application de l'équation  $Y = \sum_{k=1}^{9152} y_k$  à la population fixée, où  $y_k = 1$  si l'individu  $k$  est au chômage et est égal à 0 dans les autres cas. Pour chacun des  $R = 4,000$  échantillons, on a calculé  $Y_w$ , soit le nombre total estimatif de chômeurs  $Y_w = \sum_{k \in s} w_k y_k$ . La valeur  $\{w_k : k \in s\}$  a été déterminée au moyen des six fonctions de distance dont on a déjà parlé:

- (4) la fonction de distance IQ restrictive (IQR) ou à Logit (équation (2.9)),
- (5) la fonction de distance Huang-Fuller modifiée (HFM) ( $\alpha = 0.67$ ,  $\delta = 0.8$ ) (équation (2.10)), et
- (6) la fonction de distance à rétrécissement-minimisation (RM) ( $\alpha = 0.67$ ,  $\eta = 0.9$ ) (équation (2.11)).

Les quatre ensembles de limites qui suivent ont été appliqués aux quatre dernières fonctions de distance pour restreindre la minimisation (i)  $L = 0$ ,  $U = 4$ ; (ii)  $L = 0.4$ ,  $U = 2$ ; (iii)  $L = 0.68$ ,  $U = 1.6$  et (iv)  $L = 0.8$ ,  $U = 1.3$ . On a ainsi obtenu 18 estimateurs ponctuels. Pour chacun d'eux, le calage reposait sur les données auxiliaires issues des projections provinciales du recensement pour dix catégories d'âge/sexes mutuellement exclusives et exhaustives (catégories:  $< = 14$ , 15-24, 25-44, 45-64,  $> = 65$ , pour chaque sexe) et les quatre régions économiques de Terre-Neuve. Les données auxiliaires sur chaque individu formaient donc un vecteur de 14 unités comptant exactement deux fois la valeur un et 12 fois la valeur zéro. On a cependant dû réduire les dimensions du vecteur à 13 quand on a recouru à la méthode de Newton-Raphson pour résoudre l'équation (2.3). Par conséquent, on a postulé que  $c_k = 1$  pour les quatre premières fonctions de distance.

Nous avons calculé l'estimateur de la variance jackknife de l'équation (3.3) pour chacun des  $R = 4,000$  échantillons et des 18 estimateurs ponctuels. Nous avons aussi calculé l'estimateur de la variance de Taylor de l'équation (3.2) et procédé à la modification suggérée à la partie 3 pour les autres fonctions de distance que la fonction MCG. Notons que puisqu'on a recouru à un échantillonnage avec remise avec probabilité proportionnelle à la taille plutôt qu'à un échantillonnage sans remise au premier degré, l'estimateur de variance de l'équation (3.2) convenait tout à fait à la simulation. Enfin, on s'est servi de la formule (3.1) avec la fonction de distance MCG seulement, afin de voir ce qui se produirait si on omettait les coefficients  $g$  dans l'estimateur de la variance.

Plusieurs propriétés fréquentistes ont été examinées pour chacune des six fonctions de distance précitées. Les voici.

- (A) On a estimé le biais relatif en pour cent du nombre estimatif de chômeurs (par rapport à l'ensemble de la population) au moyen de l'équation:

$$(4.1) \quad \frac{E_M(Y_w) - Y}{Y} * 100$$

où

$$E_M(Y_w) = \frac{1}{R} \sum_{r=1}^R Y_{w_r}$$

- représente l'espérance de Monte Carlo de l'estimateur ponctuel  $Y_w$  pour  $R$  échantillons et  $Y_{w_r}$ , la valeur de  $Y_w$  pour l'échantillon  $r$ .
- (B) Le biais relatif en pour cent de l'estimateur de la variance de Taylor/jackknife (par rapport à la variance réelle) est estimé au moyen de l'équation



distance MCG) avec un plan d'échantillonnage stratifié à degrés multiples :

$$\hat{V}_T^*(Y^{w(\text{GREG})}) = \left[ \sum_{h=1}^L \frac{n_h}{n_h} - 1 \sum_{i=1}^{k_{\text{ES}} h} a_{hik} e_{hik} - \frac{1}{n_h} \sum_{i=1}^{k_{\text{ES}} h} \sum_{k \in s_h} a_{hik} e_{hik} \right]^2 \quad (3.1)$$

où  $s_h$  est l'échantillon d'individus de la  $i$ -ième unité primaire d'échantillonnage (UPÉ) et de la  $h$ -ième strate,  $a_{hik}$  est le poids d'échantillonnage original du plan d'échantillonnage pour l'individu  $k$  de l'UPÉ  $i$  et de la strate  $h$ , et  $n_h$  est le nombre d'UPÉ échantillonnées dans la strate  $h$ . De plus,  $e_{hik} = y_{hik} - x_{hik}'\beta$  donne la valeur résiduelle associée à l'estimateur de régression, où  $\beta = (\sum_{hik \in s_h} a_{hik} x_{hik}'/c_{hik})^{-1}$ . La formule «avec remise» donnée en (3.1) suréstime la variance réelle de nombreux plans d'échantillonnage (lire Särndal, Swensson et Wretman (1992, partie 4.6)). Notons cependant que sur le plan technique, cet estimateur simplifié de la variance ne correspond *pas* à l'estimateur de Taylor, même si on l'appelle souvent ainsi pour des raisons historiques. C'est pourquoi nous en ferons autant.

Hidiroglou, Fuller et Hickman (1980) proposent une amélioration à l'équation (3.1) où le coefficient  $g$  est inclus à la formule de la variance (rappelons que  $w_{hik} = a_{hik}g_{hik}$ ). Le résultat est le suivant :

$$\hat{V}_T^*(Y^{w(\text{GREG})}) = \left[ \sum_{h=1}^L \frac{n_h}{n_h} - 1 \sum_{i=1}^{k_{\text{ES}} h} w_{hik} e_{hik} - \frac{1}{n_h} \sum_{i=1}^{k_{\text{ES}} h} \sum_{k \in s_h} w_{hik} e_{hik} \right]^2 \quad (3.2)$$

Särndal (1982) suggère une équation semblable à l'équation (3.2) pour un échantillonnage à deux degrés, mais avec des estimateurs de la variance de type Yates-Grundy. De leur côté, Deville et Särndal (1992) montrent qu'une fonction de distance qui obéit à un jeu de conditions générales donnera un estimateur asymptotiquement équivalent à celui obtenu grâce à la fonction de distance MCG, c'est-à-dire  $Y^{w(\text{GREG})}$ , indiqué en (2.5). Singh et Mohl (1996) vont un peu plus loin et y intègrent la fonction de distance Huang-Fuller modifiée et la fonction de distance à rééchantillonnage. La variance asymptotique de l'estimateur de calage  $Y^w$  peut donc être considérée comme étant approximativement égale à l'estimateur  $Y^{w(\text{GREG})}$ . Cette dernière observation nous amène à une méthode permettant d'estimer la variance de Taylor commune à tous les estimateurs de calage, c'est-à-dire d'estimer la variance de  $Y^w$  en modifiant l'estimateur de variance de Taylor utilisé pour  $Y^{w(\text{GREG})}$  plutôt qu'en effectuant une nouvelle dérivation de la formule de Taylor pour chaque fonction de distance. Pour obtenir l'estimateur de la variance d'une autre fonction de distance que la fonction MCG, on peut donc recourir à l'équation (3.2) et remplacer les poids finals  $\{w_{hik}\}$  de la fonction de distance MCG par ceux de la fonction de distance à laquelle on s'intéresse.

## 4. SIMULATION DE MONTE CARLO

### 4.1 Description

Pour comparer la performance des estimateurs de calage à celle des estimateurs de la variance de Taylor et jackknife correspondants, nous avons entrepris une simulation de Monte Carlo dans laquelle nous avons examiné les propriétés fréquentielles de leur échantillon tiré d'une population finie. Les données de l'Enquête sur la population active (EPA) de décembre 1990 pour Terre-Neuve nous ont servi à créer une population finie d'où a été prélevé de façon répétitive un échantillon. L'EPA est l'enquête-ménage par échantillonnage permanente la plus importante entreprise par Statistique Canada. Les données mensuelles sur le marché du travail sont recueillies dans le cadre d'un plan d'échantillonnage complexe à degrés multiples qui comporte plusieurs niveaux de stratification. On trouvera une description détaillée du plan d'échantillonnage de cette enquête avant sa restructuration

Pour trouver l'estimateur de la variance de  $Y^w$ , sans tenir compte de la fonction de distance servant à établir les poids finals calés, on recourt couramment à la méthode du jackknife. Voici la formule de la variance applicable à un plan d'échantillonnage stratifié à degrés multiples avec remise au premier degré :

$$\hat{V}_J(Y^w) = \sum_{h=1}^L \frac{n_h}{n_h - 1} \sum_{i=1}^{k_{\text{ES}} h} (Y^w(hi) - Y^w)^2 \quad (3.3)$$

calage ne le sont qu'à la convergence. Or, il vaut souvent la peine de planifier le nombre d'itérations nécessaires pour atteindre la convergence; on peut intégrer cette spécification à l'algorithme d'itération pour faciliter le calcul. Quand la limite supérieure est franchie à cause d'une convergence trop lente, l'application de l'algorithme prend fin prématurément. Quoi qu'il en soit, les contraintes de calage sont respectées avec la fonction de distance Huang-Fuller modifiée et sa fonction de distance à rétrécissement-minimisation. Par ailleurs, avec la fonction de distance MCG restrictive et la fonction de distance IQ restrictive ce sont les contraintes relatives à la fourchette de valeurs qu'on respecte.

Le comportement des coefficients  $g$  de certaines fonctions de distance a été abondamment étudié. Il suffit de lire, par exemple, Deville, Särndal et Sautory (1993). Stukel et Boyer (1992) montrent de façon empirique que les fonctions de distance MCG et IQ, et leurs variantes restrictives, dotées de larges limites, donnent des coefficients  $g$  dont la distribution suit assez bien la normalité pour un ensemble de données particulières. Lorsqu'on impose des limites plus étroites aux fonctions de distance restrictives, la distribution révèle cependant un «empiètement» de coefficients  $g$  aux limites inférieures et supérieures. Malgré tout, les fonctions de distance restrictives semblent déboucher sur des estimations ponctuelles voisines de celles que donnent les fonctions de distance non restrictives, même avec une application très rigoureuse des limites, comme le montreront les résultats de notre étude empirique. Nous n'avons toutefois pas examiné le biais des estimateurs ponctuels et des estimateurs de variance quand on s'adresse à l'extrême des fonctions de distance. Notre analyse présente de l'intérêt pour les enquêtes comme l'EPA, dont le régime d'estimations existant a été élargi et qui permettent désormais à l'utilisateur de choisir entre les fonctions de distance MCG restrictive ou une fonction de distance à rétrécissement-minimisation, en plus de la fonction de distance MCG offerte jusqu'à présent.

### 3. ESTIMATION DE LA VARIANCE DES ESTIMATEURS DE CALAGE

La variance exacte de l'estimateur de calage  $\hat{Y}_w$  est impossible à calculer, car l'estimateur ponctuel lui-même n'est pas linéaire. De plus, il n'existe pas de méthode non biaisée explicite pour estimer la variance. C'est pourquoi on recourt souvent à des méthodes à peu près non biaisées comme celle de Taylor et celle du jackknife, dans la pratique. On se sert rarement de l'échantillonnage à tirage «avec remise» avec les plans d'échantillonnage stratifiés à degrés multiples en pratique, car on ne désire guère prélever la même unité plus d'une fois. Par conséquent, la majorité des enquêtes recourent à l'échantillonnage «sans remise», du moins lors de la première phase de l'échantillonnage. Malgré cela, si la traction de la première phase de l'échantillonnage est faible (moins de 10%, par exemple), il pourrait s'avérer raisonnable d'utiliser une formule simplifiée, supposant un échantillonnage «avec remise», à la première étape, pour estimer la variance. Pareille simplification donne l'équation qui suit pour l'estimateur de régression généralisé (fonction de

l'algorithme d'itération utilisé pour parvenir à une solution. Singh et Mohi (1996) ont testé de façon empirique diverses valeurs pour ces paramètres à partir de vastes ensembles de données. Selon eux, les valeurs  $\alpha = .67$  et  $\delta = .8$  donnent de bons résultats dans la pratique. Enfin, le coefficient  $g$  de chaque itération est

$$g_{(v-1)}^k = \frac{1 + (X - X_{(v-2)}^w)' \left( \sum_{j=1}^J a_j q_j^{(v-2)*} x_j x_j' \right)^{-1} x_k}{\sum_{k \in S} w_k^{(v-2)} x_k; v = 2, 3, \dots; \text{ où } w_k^{(v-2)} = a_k g_{(v-2)}^k, v = 2, 3, \dots}$$

où  $X_{(v-2)}^w = \sum_{k \in S} w_k^{(v-2)} x_k; v = 2, 3, \dots$ ; et où les valeurs de départ sont données par  $g_{(0)}^k = 1$  et  $w_{(0)}^k = a_k$ .

Ces deux auteurs proposent aussi une nouvelle fonction de distance qui change d'itération en itération, baptisée fonction de distance à rétrécissement-minimisation (RM). Ils montrent que l'estimateur issu de cette fonction est lui aussi asymptotiquement équivalent à l'estimateur de régression. On l'explique par:

$$F^*(w_{(v-1)}^k, a_k) = F_{SM}^*(w_{(v-1)}^k, a_k)$$

$$= (w_{(v-1)}^k)^{a_k} - (w_{(v-1)}^k)^{2/a_k}; v = 1, 2, \dots \quad (2.11)$$

où

$$a_{(v-1)}^k = \begin{cases} L^k a_k & \text{si } w_{(v-1)}^k < L^k a_k \\ U^k a_k & \text{si } w_{(v-1)}^k > U^k a_k \\ w_{(v-1)}^k & \text{autrement.} \end{cases} \quad v = 2, 3, \dots$$

Les termes des équations qui précèdent se définissent comme suit:  $L^k = \alpha L + (1 - \alpha)$ ,  $U^k = \alpha U + (1 - \alpha)$ ,  $L'' = \eta L + (1 - \eta)$  et  $U'' = \eta U + (1 - \eta)$  pour  $\alpha$  et  $\eta$ , choisis arbitrairement respectant  $0 < \alpha < \eta \leq 1$ . Comme c'est le cas précédemment, les paramètres  $\alpha$  et  $\eta$  accélèrent la convergence de l'algorithme d'itération permettant de parvenir à une solution; Singh et Mohi (1996) pensent que les valeurs  $\alpha = .67$  et  $\eta = .9$  donnent de bons résultats, en pratique. Enfin,  $w_{(v-1)}^k = a_k g_{(v-1)}^k$ ,  $v = 2, 3, \dots$  où

$$g_{(v-1)}^k = \frac{a_k}{a_{(v-2)}^k} \left[ 1 + (X - X_{(v-2)}^w)' \left( \sum_{j=1}^J a_j q_j^{(v-2)*} x_j x_j' \right)^{-1} x_k \right]; v = 2, 3, \dots$$

et où  $X_{(v-2)}^w$  est défini comme précédemment. Les valeurs de départ sont données par  $a_{(0)}^k = a_k$  et  $w_{(0)}^k = a_k$ . Une propriété de la fonction de distance Huang-Fuller modifiée et de la fonction de distance à rétrécissement-minimisation est que les contraintes de calage (équation (2.2)) sont respectées à chaque itération, tandis que les restrictions applicables à la fourchette de valeurs du coefficient  $g$  ne le sont qu'à la convergence. Avec la fonction de distance MCG restrictive et la fonction de distance IQ restrictive, les contraintes relatives à la fourchette de valeurs du coefficient  $g$  sont respectées à chaque itération, mais les restrictions de



$$Y^{w(\text{GREG})}_a = Y^a + (X - X^a)' \beta \quad (2.5)$$

$$\beta = \left( \sum_{k \in s} a_k x_k x_k' / c_k \right)^{-1} \sum_{k \in s} a_k x_k y_k / c_k \quad (2.6)$$

L'estimateur de régression correspond donc à l'estimateur HT auquel s'ajoute un terme de correction. L'inconvénient de la fonction de distance MCG est qu'elle peut donner des poids négatifs, surtout si le système fait l'objet de contraintes trop nombreuses. Dans la pratique, les poids négatifs sont rares; cependant, il est préférable de les éliminer totalement car les interpréter pourrait s'avérer difficile.

La fonction de distance par itération du quotient (IQ) se définit comme suit:

$$F^*(w_k, a_k) = F_{RR}^*(w_k, a_k) \\ = c_k [w_k \log(w_k / a_k) - w_k + a_k] \\ = a_k c_k [(w_k / a_k) \log(w_k / a_k) - (w_k / a_k) + 1]. \quad (2.7)$$

On peut montrer que calculer les coefficients  $g$  au moyen de la fonction de distance IQ et des restrictions définies en (2.3) équivaut à utiliser l'algorithme d'ajustement proportionnel itératif (API) de Deming et Stephan (1940) afin de caler les valeurs marginales connues des tableaux de fréquence à deux dimensions ou plus. À l'inverse de la fonction de distance MCG, dont la solution est fermée, les équations de calage de la fonction de distance IQ ne peuvent être résolues que par itération. Divers logiciels le permettent; le logiciel CALMAR (lire Deville, Särndal et Sautory 1993), notamment, résout les équations de calage pour la fonction de distance IQ par la méthode de Newton-Raphson plutôt qu'au moyen de l'algorithme API que proposent au départ Deming et Stephan. La fonction de distance IQ donne toujours des poids positifs; néanmoins, elle a pour défaut d'aboutir à des poids calés excessifs dans certains cas.

Des poids négatifs comme ceux de la fonction de distance MCG et des poids positifs importants comme ceux de la fonction de distance IQ ne nous intéressent pas. On peut cependant définir une fonction de distance restrictive en vertu de laquelle les poids résultants  $w_k$  seront bornés. Il suffit d'imposer des restrictions à la fonction de distance  $F(w_k / a_k)$ , de sorte que les coefficients  $g_k = w_k / a_k$  se retrouvent dans une fourchette préétablie. Pour cela, on fixe une limite inférieure  $L$  et une limite supérieure  $U$  en vertu desquelles  $L < 1 < U$ . Pour obtenir des poids positifs, on prend  $L > 0$ . Deville et Särndal (1992) présentent des variantes restrictives des deux fonctions de distance précitées, à savoir la fonction de distance MCG restrictive (MCGR) et la fonction de distance IQ restrictive (IQR) ou à logit. Huang et Fuller (1978) et Singh et Mohl (1996) proposent deux autres méthodes qui limitent les poids finals. Ces quatre fonctions de distance restrictives sont examinées plus loin. Singh et Mohl (1996) les analysent aussi en détail, mais sous un angle différent.

La fonction de distance MCG restrictive se définit ainsi:

$$F^*(w_k, a_k) = \begin{cases} c_k (w_k - a_k)^2 / a_k & \text{si } L < w_k / a_k < U \\ F_{\text{RGLS}}^*(w_k, a_k) & \text{si } L < w_k / a_k < U \\ \infty & \text{autrement.} \end{cases} \quad (2.8)$$

La fonction de distance IQ restrictive (ou à logit) s'écrit

$$F^*(w_k, a_k) = F_{\text{RRR}}^*(w_k, a_k) = \begin{cases} A^{-1} c_k [(w_k / a_k - L) \log[(w_k / a_k - L) / (1 - L)] + (U - w_k / a_k) \log[(U - w_k / a_k) / (U - 1)]] & \text{si } L < w_k / a_k < U \\ \infty & \text{autrement} \end{cases} \quad (2.9)$$

où  $A = (U - L) / [(1 - L)(U - 1)]$ . La spécification  $L = 0$ ,  $U = \infty$  donne la fonction de distance IQ. On voit aisément que ces deux fonctions de distance partagent la propriété que les poids  $w_k$  correspondants satisfont à  $L < w_k / a_k < U$ .

Huang et Fuller (1978) suggèrent une méthode pour corriger les poids de régression afin que les contraintes de calage de l'équation (2.2) soient satisfaites et que les coefficients  $g$  restent voisins de un. Singh et Mohl (1996) montrent que leur méthode correspond à la minimisation d'une fonction de distance qui change d'une itération à l'autre. Ces deux auteurs modifient aussi la méthode originale afin de permettre l'application de restrictions arbitraires aux coefficients  $g$ , restrictions semblables à celles des fonctions de distance qui précèdent, et montrent que l'estimateur résultant est asymptotiquement équivalent à l'estimateur de régression. La fonction de distance modifiée de Singh et Mohl (HFM) s'écrit ainsi:

$$F^*(w_k, a_k) = F_{\text{HFM}}^*(w_k, a_k) \\ = (w_k^{(v)} / a_k^{(v)})^{q_k} - a_k^{(v)} (w_k^{(v)} / a_k^{(v)})^{q_k} q_k \\ \text{où } q_k^{(v-1)*} = q_k^{(v-1)} q_k^{(0)} \dots q_k^{(v-1)} q_k^{(0)}, \text{ et où } v \text{ est le nombre d'itérations. Dans ce cas,} \quad (2.10)$$

$$q_k^{(v-1)} = \begin{cases} 1 & \text{si } q_k^{(v-1)} > .5 \\ 1 - \delta (q_k^{(v-1)})^2 & \text{si } .5 \leq q_k^{(v-1)} < 1 \\ (1 - \delta/4) q_k^{(v-1)} & \text{si } q_k^{(v-1)} \leq 1 \end{cases}$$

pour  $\delta$ , choisi arbitrairement afin que  $0 < \delta < 1$ . Par ailleurs,

$$z_k^{(v-1)} = \begin{cases} (g_k^{(v-1)})^{(v-1)} & \text{si } g_k^{(v-1)} \leq 1 \\ (g_k^{(v-1)})^{(v-1)} / (U' - 1) & \text{autrement} \end{cases}$$

où  $L' = \alpha L + 1 - \alpha$  et  $U' = \alpha U + 1 - \alpha$  pour  $\alpha$ , choisi arbitrairement afin que  $0 < \alpha < 1$ , et où  $L$  et  $U$  ont une valeur identique à celles des fonctions de distance restrictives qui précèdent. Les paramètres  $\alpha$  et  $\delta$  accélèrent la convergence de

## 2. FONCTIONS DE DISTANCE ET ESTIMATEURS DE CALAGE

Commençons par introduire l'idée fondamentale à la base

de l'estimation de calage. Soit  $U = \{1, \dots, k, \dots, N\}$ , l'ensemble de  $N$  unités d'une population finie d'unités. Dans les enquêtes par échantillonnage, on désire souvent estimer les paramètres d'une population finie, par exemple les totales, les moyennes et les ratios. Pour plus de simplicité, nous ne nous intéresserons qu'aux totales, mais les idées exposées ici peuvent aisément s'étendre à d'autres paramètres. Supposons donc que l'objectif consiste à estimer le total de la population  $Y = \sum_{k \in U} y_k$ , où  $y_k$  représente la valeur de  $y$ , la variable à laquelle on s'intéresse pour la  $k$ -ième unité de la population.

On prélève au hasard  $s$  unités de  $U$  selon un plan d'échantillonnage établi induisant les probabilités d'inclusion  $\pi_k = P(k \in s)$ . On suppose que ces dernières sont connues et positives. Soit  $a_k = 1/\pi_k$ , le poids d'échantillonnage associé à la  $k$ -ième unité. Enfin, supposons que les données auxiliaires prennent la forme des chiffres de population connus, pour une ou plusieurs variables auxiliaires.

Un estimateur élémentaire de  $Y$  est l'estimateur de Horvitz-Thompson (HT)

$$\hat{Y}_a = \sum_{k \in s} a_k y_k$$

L'estimateur HT ne peut intégrer les données auxiliaires qu'au stade du plan d'échantillonnage, mais il ne le fait pas nécessairement (cela dépend, on préférerait un autre estimateur en mesure d'intégrer les données auxiliaires au stade de l'estimation également. L'incorporation des données auxiliaires peut créer de nouveaux poids, notés  $w_k$ ,  $k \in s$ . Par conséquent, le nouvel estimateur s'écrit:

$$\hat{Y}_w = \sum_{k \in s} w_k y_k \quad (2.1)$$

L'approche retenue par Deville et Särndal (1992) ainsi que par Deville, Särndal et Sautory (1993) comporte la détermination des nouveaux poids  $\{w_k; k \in s\}$  par leur rapprochement au maximum des poids d'échantillonnage originaux  $\{a_k; k \in s\}$ , au moyen d'une fonction de distance spécifiée. Les restrictions imposées aux nouveaux poids font en sorte qu'une fois appliquées à chacune des variables auxiliaires, ces derniers reproduiront le chiffre de population  $X$  connu. À savoir,

$$\sum_{k \in s} w_k x_k = X \quad (2.2)$$

est valable, ce qui entraîne un problème de minimisation sous contrainte. Ici,  $x_k' = (x_{1k}, x_{2k}, \dots, x_{pk})$  représente un vecteur de longueur  $p$  incluant les valeurs des variables auxiliaires pour le  $k$ -ième individu, et les données auxiliaires d'une source extérieure sont résumées par le total vectoriel  $X = \sum_{k \in U} x_k$ .

$F^*(w_k, a_k)$  exprime la distance entre  $w_k$  et  $a_k$ . Deville et Särndal (1992) se bornent à discuter des fonctions de distance du genre  $F^*(w_k, a_k) = a_k c_k F(w_k/a_k)$  où  $w_k/a_k = g_k$ , quotient du

est minimisée par rapport à  $w_k$  où  $\lambda$  est un vecteur  $p$  de multiplicateurs de Lagrange. En effectuant la différenciation par rapport à  $w_k$ , on obtient les poids calés  $w_k = a_k g_k = a_k g(\lambda' x_k / c_k)$ , où  $g$  représente la fonction inverse de  $f$  et  $f(z) = df(z)/dz$ . Pour calculer  $w_k$ , il faut d'abord résoudre l'équation de calage qu'implique (2.2) afin d'obtenir  $\lambda$ , c'est-à-dire

$$\sum_{k \in s} a_k c_k F(w_k/a_k) - \lambda' \left( \sum_{k \in s} w_k x_k - X \right)$$

(2.2). En d'autres termes, l'expression  $\sum_{k \in s} a_k c_k F(w_k/a_k)$  est minimisée sous réserve de la contrainte Deville, Särndal et Sautory (1993). La distance totale de sorte que  $w_k = a_k$  représente un minimum local. (Lire de sorte que  $w_k = a_k$  représente un minimum local. (Lire continue, binnivoque et en outre que,  $F''(1) = 0$  et  $F'''(1) > 0$ , distance entre les poids est nulle. D'autre part,  $F'$  doit être convexe, et que  $F(1) = 0$ , si bien que lorsque  $w_k = a_k$ , la

$$\hat{Y}_w = \sum_{k \in s} a_k g_k y_k$$

s'écrit de la façon suivante:

à la pondération  $c_k = 1$ . Notons que l'équation (2.1) peut aussi être écrite de la façon suivante:  $\hat{Y}_w = \sum_{k \in s} a_k g_k y_k$

Cette solution d'un (éventuel) système non linéaire de  $p$  équations à  $p$  inconnues pourrait nécessiter le recours à une méthode d'itération quelconque, celle de Newton-Raphson, par exemple. Deville et Särndal (1992), Huang et Fuller (1978) ainsi que Singh et Mohl (1996) examinent plusieurs fonctions de distance. Les deux principales auxquelles nous nous attardons sont la fonction de distance des moindres carrés généralisés (MCG) et la fonction de distance par itération du quotient (IQ) mentionnées par Deville et Särndal (1992).

La fonction de distance MCG se définit comme suit:

$$F^*(w_k, a_k) = F_{GLS}^*(w_k, a_k)$$

$$= c_k (w_k - a_k)^2 / a_k = a_k c_k (w_k/a_k - 1)^2 \quad (2.4)$$

Elle permet d'obtenir l'estimateur de régression généralisé (ERG) bien connu dont l'estimateur par ratio, l'estimateur de régression simple et l'estimateur de stratification a posteriori simple sont des cas particuliers. De (2.3), il s'ensuit que les poids calés représentant la fonction de distance MCG sont:

$$w_k = a_k g_k = a_k [1 + (X - \hat{X})' \left( \sum_{j \in s} a_j x_j x_j' / c_j \right)^{-1} x_k / c_k]$$

où  $\hat{X}_a = \sum_{k \in s} a_k x_k$  représente l'estimateur HT de  $X$ . On peut écrire l'estimateur de  $Y$  correspondant sous la forme habituelle de l'estimateur de régression



# Estimation de la variance des estimateurs de calage: comparaison des méthodes du jackknife et de la linéarisation de Taylor

DIANA M. STUKEL, MICHAEL A. HIDIROGLOU et CARL-ERIK SÄRMNDAL<sup>1</sup>

## RÉSUMÉ

L'utilisation de données auxiliaires dans les méthodes d'estimation des enquêtes complexes, notamment l'Enquête sur la population active de Statistique Canada, ne cesse de se perfectionner. L'estimation par régression et l'estimation par itération du quotient étaient naguère les plus courantes pour intégrer les données auxiliaires à l'estimation. Il arrivait toutefois que les poids associés à l'estimateur soient négatifs ou hautement positifs. Les progrès théoriques réalisés récemment par Deville et Särndal (1992) en vue de la construction de poids «restreints» que l'on peut assujettir à une valeur positive et à un plafond nous ont incités à étudier les propriétés des estimateurs en résultant. Nous examinons ici les propriétés de diverses méthodes servant à engendrer des poids de ce genre et la variance estimative correspondante. Nous nous intéresserons en particulier à deux méthodes d'estimation de la variance en recourant à une simulation de Monte Carlo articulée sur les données de l'Enquête sur la population active. Il s'agit en l'occurrence des méthodes du jackknife et de la linéarisation de Taylor. On en conclut que les estimateurs ponctuels et les estimateurs de la variance n'entraînent qu'un biais minimal, même avec l'application de sévères «restrictions» aux poids finals.

**MOTS CLÉS:** Données auxiliaires; estimateurs par itération du quotient; estimateurs de régression; pondération restrictive.

## 1. INTRODUCTION

Les données auxiliaires comptent de nombreuses applications dans les enquêtes par échantillonnage. Un exemple typique est leur incorporation à l'estimation sous forme d'estimateurs de régression ou d'estimateurs par itération du quotient. Pour obtenir ces estimateurs, on multiplie le poids d'échantillonnage d'une unité par un facteur de correction, ce qui donne le poids final. Un inconvénient bien connu de l'estimateur de régression est que certains facteurs de correction peuvent être négatifs, si bien que le poids final le devient lui aussi. Par ailleurs, quelques facteurs de correction de l'estimateur par itération du quotient peuvent être très élevés et positifs, ce qui débouche sur des poids finals trop importants. Pour surmonter ces inconvénients, on peut recourir à une autre variété d'estimateurs, les «estimateurs de calage». Créés par Deville et Särndal (1992), ces estimateurs intègrent les données auxiliaires. Dans certains cas, on peut veiller à ce que les poids n'aient pas de valeur négative en spécifiant d'avance leurs limites inférieures et supérieures. On parvient à une «calés» en réduisant les fonctions qui mesurent l'écart entre les poids d'échantillonnage originaux et les poids finals calés, sous réserve de certaines contraintes de calage. Huang et Fuller (1978) ainsi que Singh et Mohl (1996) ont mis au point des estimateurs analogues qui conservent les propriétés précitées. Habituellement, les différences entre les estimations ponctuelles correspondant aux différentes fonctions de distance sont très faibles.

Depuis son lancement, l'Enquête sur la population active (EPA) de Statistique Canada a utilisé, à un moment ou à un

autre la technique d'estimation de la variance de Taylor aussi bien que celle du jackknife, parallèlement à l'emploi d'estimateurs de régression et d'itération du quotient. Outre l'estimateur de régression en usage jusqu'à présent, l'EPA a récemment permis l'application d'autres estimateurs de calage en vue d'éliminer le problème d'éventuels poids négatifs. Il vaut donc la peine d'examiner le comportement de ces estimateurs ponctuels et des estimateurs de la variance de Taylor et jackknife correspondants, surtout quand ils tolèrent l'imposition de limites aux poids. Tel est le sujet principal de ce document. Disons d'emblée que les méthodes de Taylor et jackknife présentent des avantages propres. La première exige considérablement moins de calculs que la seconde, mais il faut créer de nouvelles expressions pour chaque paramètre envisagé, fardéau particulièrement lourd dans le cas des enquêtes à fins multiples dans lesquelles on est susceptible de s'intéresser à un grand nombre de paramètres. De son côté, la méthode du jackknife, n'exige pas l'établissement d'expressions encombrantes de la variance pour chaque nouveau paramètre; la fonction de l'estimateur ponctuel suffit.

Dans la partie 2, on expose les principes théoriques de l'estimation par calage et introduit une série de fonctions de distance connexes; la partie 3 traite de la variance des estimateurs de calage; on trouvera à la partie 4 les résultats d'une étude de simulation de Monte Carlo qui établit le biais des estimateurs ponctuels et des estimateurs de la variance de Taylor et jackknife correspondants (par rapport à une variance «réelle») pour diverses fonctions de distance issues de la théorie de calage. La partie 5 sert de conclusion.

<sup>1</sup> Diana M. Stukel, Division des méthodes d'enquêtes des ménages, et Michael A. Hidiroglou, Division des méthodes d'enquêtes-entreprises, Statistique Canada, Ottawa, Ontario, K1A 0T6; Carl-Erik Särndal, Département de Mathématiques et de Statistique, Université de Montréal, C.P. 6128, Succursale A, Montréal, P.Q., H3C 3J7.

## BIBLIOGRAPHIE

- Singh et Mohl: Comprendre les estimateurs de calage dans les enquêtes par échantillonnage
- SÄRNDAAL, C.-E. (1980). On  $\pi$ -inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 67, 639-650.
- SINGH, A.C. (1993). On weight adjustment in survey sampling. Article discuté à la 18<sup>ième</sup> réunion du Comité consultatif des méthodes statistiques, Statistique Canada, Ottawa, Octobre 25-26.
- SINGH, A.C., et MOHL, C.A. (1997). Calibration estimators with application to FAMEX survey and computer program documentation. Documents de travail de la direction de la méthodologie, Statistique Canada.
- SINGH, M.P., DREW, J.D., GAMBINO, J.G., et MAYDA, F. (1990). *Méthodologie de l'enquête sur la Population active du Canada: 1984-1990*, N° 71-526 au catalogue, Statistique Canada.
- STUKEL, D.M., et BOYER, R. (1992). Calibration Estimation: An Application to the Canadian Labour Force Survey. Documents de travail de la direction de la méthodologie, SSMD 92-009E, Statistique Canada.
- STUKEL, D.M., HIDIROGLOU, M.A., et SÄRNDAAL, C.-E. (1996). Estimation de la variance des estimateurs de calage: comparaison des méthodes du jackknife et de la linéarisation de Taylor. *Techniques d'enquête*, 22, 117-126.
- BARDSLEY, P., et CHAMBERS, R.L. (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics*, 33, 290-299.
- DEVILLE, J.-C., et SÄRNDAAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- DEVILLE, J.-C., SÄRNDAAL, C.-E., et SAUTORY, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.
- HUANG, E.T., et FULLER, W.A. (1978). Nonnegative regression estimation for sample survey data. *Proceedings of the Social Statistics Section, American Statistical Association*, 300-305.
- FULLER, W.A., LOUGHIN, M.M., et BAKER, H.D. (1994). Production de poids de régression en situation de non-réponse et application à la Nationwide Food Consumption Survey de 1987-1988. *Techniques d'enquête*, 20, 79-89.
- LEMAÎTRE, G., et DUFOUR, J. (1987). Une méthode intégrée de pondération des personnes et des familles. *Techniques d'enquête*, 13, 211-220.



**A2. MÉTHODE 2 (MDI-u)**

La solution est obtenue par itération selon les étapes suivantes pour  $v = 1, 2, \dots$

(i) Poser le niveau de tolérance  $\delta \geq 0$  de façon à respecter

les CB pour une valeur faible quelconque.

(ii) Pour la  $v$ -ième itération, calculer  $f_k^{(v)}$ ,  $k = 1$  à  $n$ , à

partir de (1) en posant  $\Gamma^{v-1} = \text{diag}(c_k^{(v)})$ , sous la forme

(iii) Pour  $v = 1, 2, \dots$ , calculer  $g_k^{(v)}$  et ensuite  $c_k^{(v)}$  à partir de

(iv) Répéter les étapes (ii)-(iii) jusqu'à ce que les CB soient

respectées jusqu'au niveau de tolérance  $\delta$  ou jusqu'à

ce que le nombre d'itérations atteigne son maximum

$v^{\max}$ . La dernière itération donne  $c_k^{\text{MDI-u}}$ .

**A3. MÉTHODE 3 (SMCS)**

La solution est obtenue par itération comme suit.

(i) Poser les RAFV, c'est-à-dire choisir  $L$  et  $U$ ,

(ii) Poser le niveau de tolérance  $\epsilon \geq 0$  à une faible valeur

de façon à respecter les RAFV.

(iii) Choisir un paramètre  $\alpha$  entre 0 et 1 (p. ex. 2/3) et

poser  $L' = \alpha L + 1 - \alpha$ ,  $U' = \alpha U + 1 - \alpha$ . Une valeur

implicite de 1 pour  $\alpha$  est également admise et alors

(iv) Pour la  $v$ -ième itération avec  $g_k^{(0)} = 1$ , définir  $z_k^{(v-1)} =$

$(g_k^{(v-1)} - 1)/(L' - 1)$  si  $g_k^{(v-1)} \leq 1$ ;  $(g_k^{(v-1)} - 1)/(U' - 1)$

autrement.

(v) Choisir un autre paramètre  $\beta$  entre 0 et 1 (p. ex. 4/5).

Poser  $q_k^{(v-1)} = 1$  si  $z_k^{(v-1)} < 1/2$ ;  $1 - \beta(z_k^{(v-1)} - 1/2)$  si

$1/2 \leq z_k^{(v-1)} < 1$ ;  $1 - \beta/4$  si  $z_k^{(v-1)} \geq 1$ ; et alors

définir pour  $v = 1, 2, \dots$ ,  $q_k^{(v)} = q_k^{(v-1)}$  ou

$q_k^{(0)} = 1$ . Noter le jumelage des facteurs  $q$  dans la

définition de  $q_k^{(v-1)}$ .

(vi) Calculer  $f_k^{(v)}$  à partir de (1) en posant  $\Gamma^{v-1} =$

$\text{diag}(h_k q_k^{(v)} g_k^{(v)})$ , et  $\frac{x}{\Gamma^{v-1}} = \frac{x}{\Gamma^{(0)}}$  pour tous les  $v$ .

(vii) Trouver  $g_k^{(v)}$  sous la forme  $1 + q_k^{(v-1)} f_k^{(v)}$  et ensuite

$c_k^{(v)}$  sous la forme  $h_k g_k^{(v)}$ .

(viii) Répéter les étapes (vi)-(vii) jusqu'à ce que les RAFV

soient respectées jusqu'au niveau de tolérance  $\epsilon$  ou

$v = v^{\max}$ . La dernière itération donne  $c_k^{\text{SMCS}}$ . La valeur

de  $\beta$  devrait rester la même à chaque itération.

**A4. MÉTHODE 4 (SM)**

Cette méthode comporte les étapes qui suivent, exécutées

par itération.

(i)-(ii) Comme pour la méthode 3.

(iii) Choisir les paramètres  $\alpha, \eta$ ,  $0 < \alpha \leq \eta \leq 1$ , (p. ex.

$\alpha = 2/3$ ,  $\eta = 9/10$ ) et définir

$L' = \alpha L + (1 - \alpha)$ ,  $U' = \alpha U + (1 - \alpha)$

$L'' = \eta L + (1 - \eta)$ ,  $U'' = \eta U + (1 - \eta)$ .

L'option implicite pour  $\alpha$  et  $\eta$  est 1 et alors

$L' = L'' = L$ ,  $U' = U'' = U$ .

(Rétrecissement). Le  $c_k^{(v)}$  de la  $v$ -ième itération est

rétréci de façon à donner  $c_k^{(v)*}$  conformément à

$c_k^{(v)*} = L' h_k^{(v)} c_k^{(v)}$  si  $c_k^{(v)} > U'' h_k^{(v)}$  si  $c_k^{(v)} > U'' h_k^{(v)}$ ;  $c_k^{(v)}$

autrement. Pour  $v = 0$ ,  $c_k^{(0)*} = c_k^{(0)} = h_k^{(v)}$ .

**REMERCIEMENTS**

Les auteurs tiennent à remercier C.-E. Sarnadal pour des discussions fructueuses et U. Nevrbaumont d'avoir fourni les données FAMBX. Les auteurs remercient également les rapporteurs de leurs remarques très utiles. Les recherches du premier auteur ont été financées en partie par une subvention du CRSNG accordée à l'Université Carleton à Ottawa.

**A7. MÉTHODE 7 (GMDI)**

L'algorithme itératif comporte les étapes suivantes.

(i)-(ii) Comme pour la méthode 5.

(iii) Calculer  $f_k^{(v)}$  à partir de (1) en posant  $\Gamma^{v-1} =$

$\text{diag}(h_k d_k^{(v-1)})$  où  $d_k^{(v-1)}$  est analogue à  $d_k^{(v)}$  de la

section 2.7.

(iv) À l'aide de  $x_k^{(v)} = x_k^{(v-1)} + f_k^{(v)}$ , trouver  $g_k^{(v)}$  à

même la formule pour  $g_k^{(v)}$  donnée à la section 2.7, et

ensuite  $c_k^{(v)}$  sous la forme  $h_k g_k^{(v)}$ .

(v) Répéter les étapes (iii)-(iv) jusqu'à ce que les CB

soient respectées au niveau de tolérance  $\delta$  ou

$v = v^{\max}$ . La dernière itération donne  $c_k^{\text{GMDI}}$ .

**A6. MÉTHODE 6 (MDI-r)**

L'algorithme itératif comporte les étapes suivantes.

(i)-(ii) Comme pour la méthode 5.

(iii) Calculer  $f_k^{(v)}$  à partir de (1) en posant  $\Gamma^{v-1} =$

$\text{diag}(c_k^{(v-1)} a_k^{(v-1)})$  où  $a_k^{(v-1)}$  se définit comme à l'étape

(iii) de la méthode 5.

(iv) Poser  $g_k^{(0)} = 1$  et calculer  $g_k^{(v)} = g_k^{(v-1)} \exp(f_k^{(v)})$  si

$L \leq g_k^{(v-1)} + f_k^{(v)}$  si  $L \leq g_k^{(v-1)}$ ; autrement tronquer  $g_k^{(v)}$  à

$L$  ou  $U$  selon le cas, et ensuite  $c_k^{(v)}$  comme  $h_k g_k^{(v)}$ .

(v) Répéter les étapes (iii)-(iv) jusqu'à ce que les CB

soient respectées au niveau de tolérance  $\delta$  ou

$v = v^{\max}$ . La dernière itération donne  $c_k^{\text{MDI-r}}$ .

**A5. MÉTHODE 5 (MCS-r)**

L'algorithme itératif comporte les étapes suivantes.

(i) Poser  $L$  et  $U$ .

(ii) Poser le niveau de tolérance  $\delta \geq 0$  de façon à respecter

les CB.

(iii) Calculer  $f_k^{(v)}$  à partir de (1) en posant  $\Gamma^{v-1} =$

$\text{diag}(h_k a_k^{(v-1)})$  où  $a_k^{(v-1)} = 1$  si  $g_k^{(v-1)}$  a été tronqué

à  $L$  ou  $U$ , et 0 autrement.

(iv) Poser  $g_k^{(0)} = 1$  et calculer  $g_k^{(v)}$  sous la forme

$g_k^{(v-1)} + f_k^{(v)}$  si  $L \leq g_k^{(v-1)}$ ; autrement tronquer  $g_k^{(v)}$  à

$L$  ou  $U$  selon le cas, et ensuite  $c_k^{(v)}$  sous la forme

$h_k g_k^{(v)}$ .

(v) Répéter les étapes (iii)-(iv) jusqu'à ce que les CB

soient respectées au niveau de tolérance  $\delta$  ou

$v = v^{\max}$ . La dernière itération donne  $c_k^{\text{MCS-r}}$ .

**Tableau 2c**  
Différence des estimations ponctuelles et de la précision liée à l'estimateur de régression ( $\alpha = .67, \beta = .8, \eta = .9, \epsilon = .01, v_{\max} = 10$ ) Saskatoon:  $L = 0.88, U = 1.12$

	Logement en propriété		Mobiliertéquipement	
	DR	PR	DR	PR
Famille I				
SMCS	-0.001	1.001	-0.001	0.999
SM	-0.000	1.001	-0.000	0.999
Famille II				
MCS-r	0.000	0.999	0.000	0.994
MDI-r	0.002	0.997	0.002	0.994
GMDI	-0.000	1.007	-0.000	0.990

Famille I				
SMCS	0.000	1.013	-0.001	0.999
SM	-0.000	1.002	-0.000	0.998
Famille II				
MCS-r	0.000	0.990	0.000	0.994
MDI-r	0.002	1.001	0.002	0.983
GMDI	0.000	0.977	-0.000	0.990

- Notes:
1. Dans le cas non restrictif (ou sans limites), les mesures correspondantes pour la méthode itérative du quotient (MDI-u) relativement à la régression sont (0.002, 1.000), (0.002, 1.002) et (0.002, 0.995) pour les quatre variables à l'étude respectivement.
  2. Durant la procédure du jackknife, la méthode SM n'a pas atteint la convergence après dix itérations pour quatre pseudorépétitions (d'un total de 94).

### 3.5.4 Nombre de calculs à effectuer

Pour Regina (tableau 1), en présence de limites larges, les limites se resserrèrent, la plupart des méthodes exigent plus d'itérations pour atteindre la convergence. Afin de vérifier à quel point les limites pouvaient être resserrées avant que des problèmes de convergence ne surviennent, nous avons utilisé trois autres séries de limites avec  $[L, U] = [0.425, 2.35]$ ,  $[0.45, 2.22]$  et  $[0.475, 2.11]$ . Ces résultats ne sont pas indiqués dans le tableau. Pour un  $v_{\max}$  de 10, la méthode SM ne converge pas pour  $[0.425, 2.35]$ . Ni la méthode SMCS ni la méthode GMDI ne convergent pour  $[0.45, 2.22]$  et les méthodes MCS-r et MDI-r posent des problèmes de convergence pour  $[0.475, 2.11]$ . Pour Saskatoon (tableau 1), étant donné les limites choisies, chaque méthode exige une ou deux itérations seulement. Pour un  $v_{\max}$  de 10, à mesure que les limites se resserrèrent jusqu'à  $[0.92, 1.08]$ , la méthode SM ne converge pas. Pour  $[0.93, 1.07]$ , les méthodes SMCS, MCS-r et MDI-r posent des problèmes de convergence et, en fin de compte, pour  $[0.96, 1.06]$ , la méthode GMDI suscite des problèmes.

## 4. DISCUSSION

Même si les résultats numériques pour quelques variables de deux domaines différents examinés dans le présent article sont trop limités pour que l'on puisse tirer des conclusions

générales, les résultats fondés sur une analyse descriptive sont tout de même intéressants et peuvent fournir certaines indications utiles en pratique. Celles-ci sont résumées dans les observations qui suivent. En présence de limites larges, toutes les méthodes restrictives semblent fonctionner presque comme la méthode de régression. Toutefois, en présence de limites étroites, il semble y avoir des différences d'estimations ponctuelles et, en particulier, de précision estimée. Cette dernière observation mérite clairement une étude plus poussée si l'on considère que toutes les méthodes sont asymptotiquement équivalentes à la méthode de régression. Une étude de simulation serait souhaitable à cet égard. La récente étude de Stukel, Hidiroglou et Särndal (1996) élucide un peu cette question. De plus, en présence de limites étroites, il est possible qu'il n'y ait pas de convergence pour le nombre indiqué d'itérations même si une solution existe. Ce problème peut devenir plus manifeste lorsqu'on traite de répétitions de type jackknife. Par conséquent, il importe d'être prudent dans le choix du nombre maximum d'itérations pour les limites étroites. Enfin, dans la pratique, il est possible que même en présence d'exigences minimales pour les CE et les RAFV, aucun des estimateurs de calage n'entraîne une convergence pour un nombre raisonnable d'itérations. Dans une telle situation, il serait intéressant de déterminer si la conformité au plan (asymptotique) des estimateurs de calage pourrait être conservée tout en admettant un écart par rapport aux CE. L'idée de Bardsley et Chambers (1984) d'avoir recours à une régression écrite, même si elle ne se situe pas dans le contexte de la conformité au plan, pourrait avoir une certaine utilité à cet égard. L'étude de ce problème se poursuit en collaboration avec J.N.K. Rao.

## ANNEXE

Nous indiquons ci-après des algorithmes de calcul pour l'ensemble des sept méthodes de correction des poids. Ces algorithmes ont servi à rédiger des programmes informatiques en logiciel GAUSS pour les exemples numériques présentés dans le présent article. Pour toutes les méthodes, une forme quelconque de l'expression ci-dessous désignée par le  $n$ -vecteur  $f^{(v)}$  est répétée pour le calcul de  $c_k^{(v)}$  pour  $v = 1, 2, \dots$ .

$$f^{(v)} \equiv X(X' T^{v-1} X)^{-1} (T^{v-1} r_x - \frac{1}{v} r_x^{(v-1)}) \quad (1)$$

où  $T^{v-1}$  est une matrice diagonale  $n \times n$  définie ci-dessous dans l'algorithme pour chaque méthode. Au début  $T_0 = \text{diag}(h) = \sum x_k h_k$ .

### A1. MÉTHODE 1 (MCS-u)

La solution, non itérative, est obtenue en deux étapes comme suit.

- (i) Calculer  $f_k^{(1)}$ ,  $k = 1$  à  $n$  à partir de (1) en posant  $T^{v-1} = T_0$ .
- (ii) Calculer  $g_k$  sous la forme  $1 + f_k^{(1)}$  et ensuite  $c_k^{\text{MCS-u}}$  sous la forme  $h_k g_k$ .



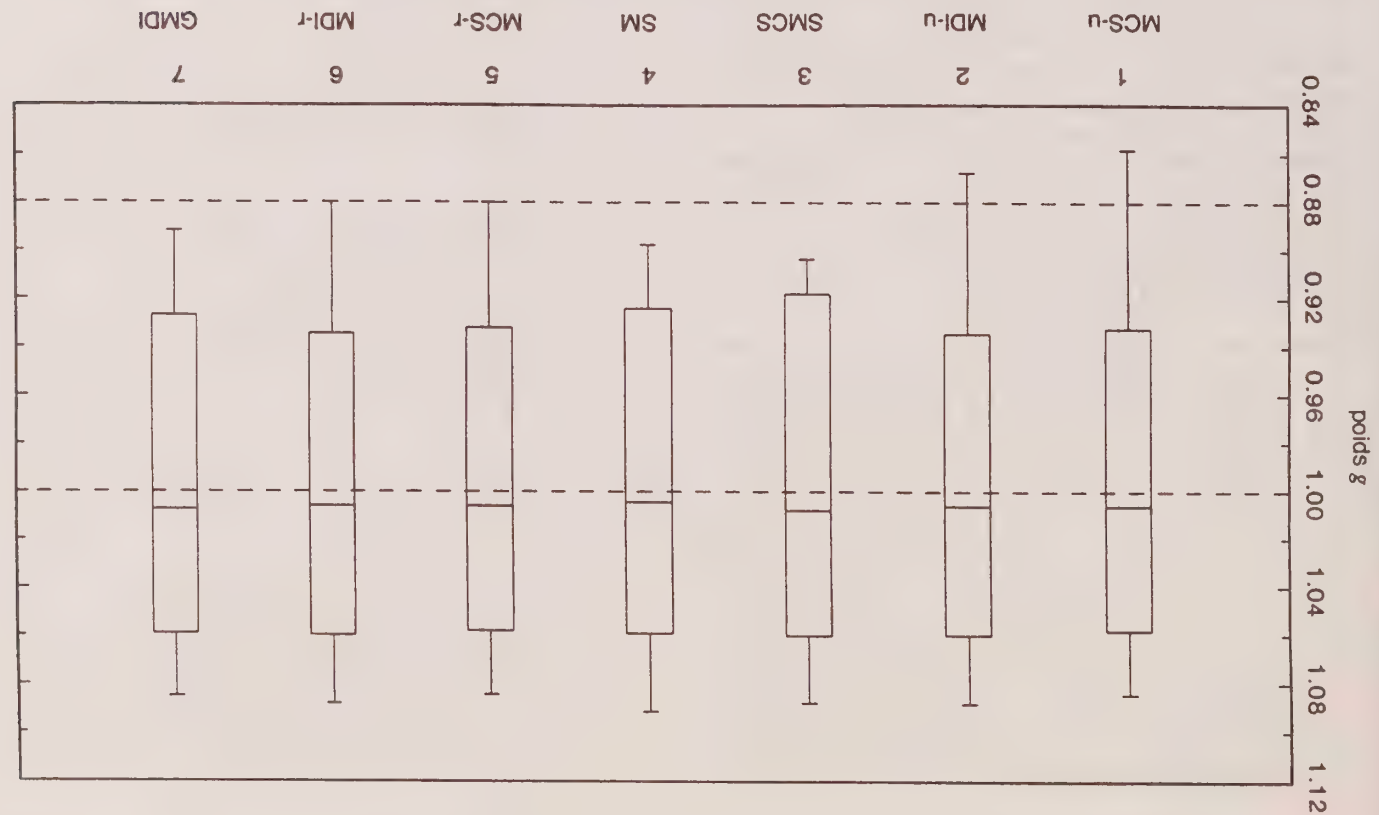


Figure 2. Tracé en boîte: poids  $g$  pour les données FAMEX de Saskatoon ( $L = 0.88$ ,  $U = 1.12$ )

Tableau 2a

Différence des estimations ponctuelles et de la précision liée à l'estimateur de régression ( $\alpha = .67$ ,  $\beta = .8$ ,  $\eta = .9$ ,  $\epsilon = \delta = .01$ ,  $v_{\max} = 10$ ) Régina:  $L = 0.2$ ,  $U = 5.0$  (Limites larges)

Logement en propriété		MobiliervÉquipement	
DR	PR	DR	PR
<b>Famille I</b>			
SMCS	-0.043	1.047	0.001
SM	-0.036	1.032	-0.002
<b>Famille II</b>			
MCS-r	-0.032	1.035	0.002
MDI-r	-0.033	0.991	-0.008
GMDI	-0.037	0.999	-0.004
<b>Vêtements pour femmes</b>			
SMCS	0.015	0.931	0.009
SM	0.010	0.951	0.006
<b>Famille I</b>			
SMCS	0.952	0.009	0.968
SM	0.968	0.008	0.964
<b>Famille II</b>			
MCS-r	0.011	0.950	0.008
MDI-r	0.007	0.911	-0.001
GMDI	0.009	0.940	0.002

Notes: 1. DR and PR désignent respectivement la «différence relative» et la «précision relative». 2. Dans le cas non restrictif (ou sans limites), les mesures correspondantes pour la méthode itérative du quotient (MDI-u) relativement sont (-0.034, 1.005), (-0.008, 1.049), (0.004, 0.968) et (0.002, 0.980) pour les quatre variables à l'étude respectivement.

Tableau 2b

Différence des estimations ponctuelles et de la précision liée à l'estimateur de régression ( $\alpha = .67$ ,  $\beta = .8$ ,  $\eta = .9$ ,  $\epsilon = \delta = .01$ ,  $v_{\max} = 10$ ) Régina:  $L = 0.4$ ,  $U = 2.5$  (Limites étroites)

Logement en propriété		MobiliervÉquipement	
DR	PR	DR	PR
<b>Famille I</b>			
SMCS	-0.056	1.100	0.012
SM	-0.055	0.992	0.017
<b>Famille II</b>			
MCS-r	-0.048	1.073	0.008
MDI-r	-0.045	1.087	0.012
GMDI	-0.047	1.077	0.009
<b>Vêtements pour femmes</b>			
SMCS	0.024	0.917	0.038
SM	0.025	0.917	0.024
<b>Famille I</b>			
SMCS	0.020	0.904	0.012
MDI-r	0.025	0.888	0.012
GMDI	0.021	0.938	0.018
<b>Vêtements pour hommes</b>			
SMCS	0.808	0.038	0.801
SM	0.808	0.038	0.801
<b>Famille I</b>			
SMCS	0.952	0.008	0.952
MDI-r	0.965	0.012	0.965
GMDI	1.006	0.009	1.006

Note: Durant la procédure du jackknife, la méthode SM n'a pas atteint la convergence après dix itérations pour quatre pseudorépétitions (d'un total de 111).

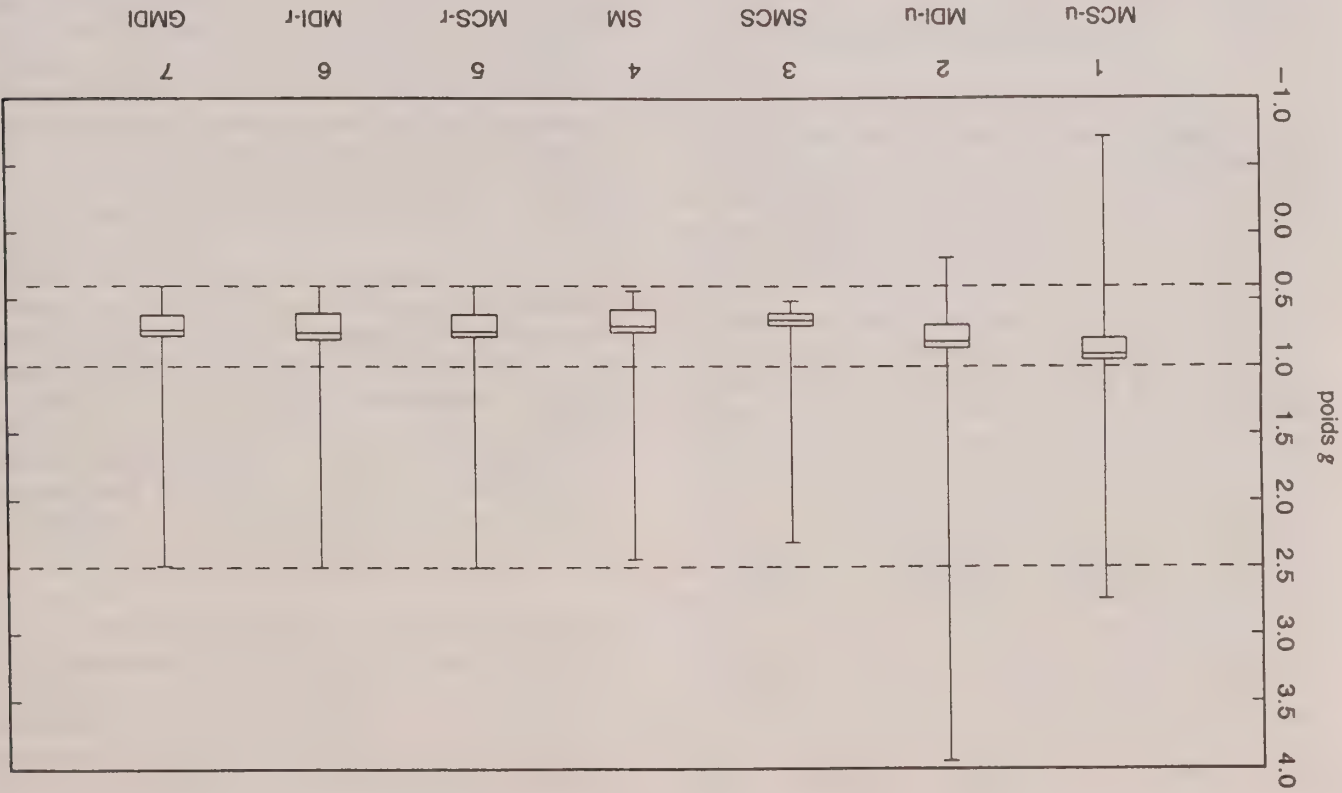


Figure 1. Tracé en boîte: poids  $g$  pour les données FAMEX de Regina ( $L = 0.4$ ,  $U = 2.5$ )

**Tableau 1**  
Nombre d'itérations et écart type SD( $g$ )  
( $\alpha = .67$ ,  $\beta = .8$ ,  $\eta = .9$ ,  $\epsilon = \delta = .01$ ,  $v_{\max} = 10$ )

Méthode	Regina		Saskatoon	
	Nombre d'itérations (Limites larges) $L = 0.2$ , $U = 5.0$	SD( $g$ ) Nombre d'itérations (Limites étroites) $L = 0.4$ , $U = 2.5$	Nombre d'itérations SD( $g$ )	Nombre d'itérations SD( $g$ )
Famille I				
SMCS	2	0.647	3	0.702
SM	2	0.636	4	0.689
Famille II				
MCS-r	2	0.628	3	0.654
MDI-r	3	0.642	3	0.660
GMDI	3	0.640	2	0.659

**Note:** Pour le cas non restrictif (ou sans limites), le nombre d'itérations et l'écart type SD( $g$ ) sont: pour les méthodes MCS-u et MDI-u de Regina on a (1,0.599) et (3,0.647) respectivement; pour les méthodes MCS-u et MDI-u de Saskatoon on a (1,0.070) et (1,0.069) respectivement.

Pour ce qui est des données de Saskatoon, on trouvera à la figure 2 un tracé en boîte de poids  $g$  où  $L = 0.88$  et  $U = 1.12$ . Pour la méthode de régression aussi bien que pour la méthode itérative du quotient, 5,6% environ se trouvent en dessous de  $L$  et 0% au-dessus de  $U$ . Toutes les méthodes comportent un intervalle interquartile semblable pour les poids  $g$ , les valeurs médianes étant légèrement supérieures à 1. On voit également

dans le tableau 1 que l'écart type SD( $g$ ) est à peu près le même et relativement faible pour toutes les méthodes (restrictives et non restrictives).

### 3.5.2 Différence relative des estimations ponctuelles

Les tableaux 2(a) et (b) indiquent que, pour Regina, en présence de limites larges, la DR est faible pour toutes les méthodes pour chacune des variables. En réalité, elle est négligeable sauf pour la variable «logement en propriété» pour laquelle elle est généralement inférieure à 4%. Toutefois, en présence de limites étroites, elle augmente dans une certaine mesure, tout en restant faible avec des valeurs variant entre 1% et 5%. Pour Saskatoon (tableau 2c), compte tenu des limites données, la DR est négligeable pour toutes les méthodes.

### 3.5.3 Précision relative prévue des estimations

Pour Regina, en présence de limites larges, la PR se situe généralement à 5% près (de la précision de l'estimateur de régression) pour toutes les méthodes et toutes les variables sauf pour la méthode MDI-r comportant la variable «vêtements pour femmes» pour laquelle elle est inférieure de 9%. Toutefois, en présence de limites étroites, la PR varie davantage et se situe généralement à 9% près sauf pour les méthodes SMCS et SM comportant la variable «vêtements pour hommes» (PR inférieure de 20%) et la méthode MDI-r comportant la variable «vêtements pour femmes» pour laquelle la PR est inférieure de 11%. Pour Saskatoon (tableau 2c), compte tenu des limites choisies, la PR est proche de 1 dans tous les cas.



L'échantillon est de 111 et 94. Le nombre total de ménages échantillonnés est de 321 et 278, le nombre correspondant de personnes ( $n$ ) étant de 797 et 712.

### 3.2 Contraintes d'établissement, restrictions applicables à la fourchette de valeurs et poids communs par ménage

Le nombre ( $p$ ) de CE est de quatre par domaine. Elles correspondent aux chiffres démographiques pour les quatre groupes: âge < 15, âge ≥ 15, ménages à une personne et ménages à deux personnes ou plus. Les chiffres correspondants sont 40696, 139047, 12746 et 48457 pour Regina, et 42544, 139299, 20628 et 52059 pour Saskatoon. Ainsi, le nombre total de ménages pour les deux domaines est de 61203 et 72687 respectivement, et la taille correspondante de la population ( $N$ ) est de 179743 et 181843. Les variables auxiliaires dans ce cas-ci sont des indicateurs pour les quatre groupes ci-dessus.

Pour ce qui est de Regina, les (min, max) des poids  $g$  correspondent à (−0.72, 2.74) et à (0.19, 3.95) respectivement pour la méthode de régression et la méthode itérative du quotient. Il est donc utile de les rendre non négatifs pour la régression et de réduire les poids élevés pour l'itération du quotient. On choisit deux types de RAFV, un qui comporte des limites assez larges avec  $L = 1/5$  et  $U = 5$ , l'autre qui comporte des limites assez étroites avec  $L = 2/5$  et  $U = 5/2$ . Pour ce qui est de Saskatoon, les (min, max) des poids  $g$  correspondent à (0.86, 1.08) et à (0.87, 1.09) respectivement pour la méthode de régression et la méthode itérative du quotient. À noter que les deux méthodes donnent des poids  $g$  proches de 1, de sorte qu'il n'y a pas de besoin réel de RAFV. Toutefois, par souci d'illustration, nous choisissons  $L = 0.88$  et  $U = 1.12$ .

Les poids d'échantillonnage initiaux ou poids  $h$  des personnes d'un même ménage sont communs et égaux au poids de ce ménage. Il est souhaitable que, après le calage, tous les membres d'un ménage aient les mêmes poids  $c$ . Il suffit pour cela de modifier la matrice  $X$  de façon que les valeurs  $x_j$  pour chaque personne du même ménage soient communes et identiques à la valeur moyenne pour le ménage (voir p. ex. Lemaître et Dufour 1987). Nous effectuons également une mise à l'échelle initiale pour les poids  $h$  de façon que leur somme soit  $N$ ; cela ressemble à la modification apportée par Häjek à l'estimateur de Horvitz-Thompson. Cette mise à l'échelle permet essentiellement de redéfinir  $[L, U]$  pour les rendre utiles dans le calage des poids  $h$ .

### 3.3 Mesures descriptives pour la comparaison

Avant de pouvoir comparer différentes méthodes, nous examinons quatre types de mesures descriptives:

- i) des données statistiques sommaires pour la distribution des poids  $g$ ;
- ii) des estimations ponctuelles pour plusieurs variables;
- iii) la précision prévue des estimations de calage;
- iv) le nombre de calculs à effectuer pour chaque méthode.

La première mesure est un résumé graphique qui utilise un tracé en boîte pour les poids  $g$  et l'écart type des poids  $g$ .

$SD(g)$ , défini par  $[N^{-1} \sum_{k=1}^n h_k^2 (g_k - \bar{g})^2]^{1/2}$ . À noter que la moyenne des poids  $g$ , c'est-à-dire  $N^{-1} \sum_{k=1}^n h_k g_k$ , est 1 étant donné que  $\sum h_k = \sum c_k = N$ , et le  $SD(g)$  égaie lui aussi  $[N^{-1} \sum_{k=1}^n (c_k - h_k)^2 / h_k]^{1/2}$ , la racine carrée d'une distance normalisée de type chi carré pour la mesure de la proximité entre les poids  $h$  et  $c$ . À des fins de comparaison des estimations ponctuelles et de leur précision pour l'évaluation du paramètre de chaque variable  $y$  d'intérêt, nous calculons la différence relative (DR) et la précision relative (PR) pour ce qui est des poids MCS-u, c'est-à-dire relativement à l'estimateur de régression. Un estimateur fondé sur des poids  $c$  étant noté estimateur  $c$ , nous avons une DR de (estimateur  $c$  moins estimateur de régression) divisée par l'estimateur de régression, et une PR de ET (erreur-type) divisée par ET (estimateur  $c$ ). À noter que, pour les exemples numériques à l'étude, les variances sont calculées à l'aide de la méthode du jackknife en éliminant les UPB. Enfin, le nombre de calculs à effectuer est exprimé en fonction du nombre d'itérations. Des tests ont montré que, pour toutes les méthodes restrictives, chaque itération exige à peu près le même temps, de sorte que le nombre d'itérations requis pour la convergence représente un bon critère de comparaison du nombre de calculs à effectuer.

### 3.4 Description d'autres paramètres

Nous devons également décrire d'autres paramètres, notamment  $\alpha$ ,  $\beta$  pour SMCS et  $\alpha$ ,  $\eta$  pour SM. De façon empirique, les valeurs de  $\alpha = 0.67$ ,  $\eta = 0.9$  et  $\beta = 0.8$  fonctionnent bien. On pose les niveaux de tolérance  $\epsilon$  pour la famille I et  $\delta$  pour la famille II à 0.01, et  $v_{\max}$  à 10.

### 3.5 Résultats: Analyse descriptive

#### 3.5.1 Distribution des poids $g$

Considérons d'abord les données pour Regina. On trouvera à la figure 1 un tracé en boîte de la distribution des poids  $g$  avec  $L = 0.4$  et  $U = 2.5$ . À noter qu'il y a des poids  $g$  négatifs (et donc des poids  $c$  négatifs) pour la méthode MCS-u et des poids  $g$  larges (produisant des poids  $c$  larges) pour la méthode MDI-u. Pour la méthode MCS-u, la fraction inférieure de poids  $g$  inférieurs à 0 est de 4.9% et la fraction inférieure à 0.4 est de 5.9%, la fraction supérieure à 2.5 étant de 1.25% et celle supérieure à 3.5 de 0%. Pour les méthodes MDI-u, la fraction inférieure à 0.4 est de 4.9% et la fraction supérieure à 2.5 est de 4.3%, la fraction supérieure à 3.5 étant de 1.25%. Ainsi, les deux méthodes produisent des poids  $c$  qui dépassent la limite pour ce qui est des RAFV à limites étroites. Les méthodes restrictives du point de vue de la fourchette de valeurs comportent toutes des poids  $g$  médians situés entre 0.65 et 0.75; toutefois, les poids  $g$  de la méthode SMCS offrent le plus de grappes autour de la médiane. Le tableau 1 indique que, en présence de limites larges, l'écart-type  $SD(g)$  pour chaque méthode restrictive est légèrement supérieur (de 7% environ) relativement à la méthode de régression, tandis que pour des limites étroites, la différence augmente jusqu'à 15% environ pour la famille I et à 10% environ pour la famille II.



minimisation de la fonction de distance  $\Delta_{MDI-r}(c, h) = \sum_{k=1}^n [c_k \log(c_k/h_k) - c_k + h_k]$  si  $Lh_k \leq c_k \leq Uh_k$ ;  $\infty$  autrement, sous réserve des CE. À noter que dans la pratique, les facteurs de troncature ne sont pas nécessaires, séparément, pour le calcul de  $g_k^{(v)}$ . On trouvera à l'annexe A6 l'algorithme de calcul pour MDI-r.

## 2.7 MÉTHODE 7 (logit ou information de

discrimination modifiée généralisée, GMDI)

C'est la dernière méthode examinée. Cette méthode

bien connue de la famille II provient de Deville et Sæmdal (1992). Comme pour la méthode itérative du quotient, nous commençons d'abord avec  $\exp(x_k^* \lambda)$  et une transformation inverse de type logit nous permet de veiller à ce que le coefficient de correction respecte les RAFV. Le modèle de coefficient de correction est donné par  $g_k = [(U - 1) + (1 - L) \exp(Ax_k^* \lambda)]^{-1} [L(U - 1) + U(1 - L) \exp(Ax_k^* \lambda)]$ , où  $A = (1 - L)^{-1}(U - 1)^{-1}(U - L)$ . Ce coefficient de correction, contrairement à d'autres méthodes, se trouve obliquement à l'intérieur de l'intervalle  $[L, U]$ , c'est-à-dire qu'il n'a pas des valeurs de limite. Lorsque  $L \rightarrow 0$  et  $U \rightarrow \infty$ , le facteur est réduit à la forme logit inverse familière,  $\exp(x_k^* \lambda) / [1 + \exp(x_k^* \lambda)]$ . On obtient le paramètre modèle  $\lambda$  par itération de façon à respecter les CE. En commençant par  $\lambda_{MCS-n}^{(1)}$  sous la forme  $\lambda_{(1)}$  pour l'itération 1, nous corrigeons selon un terme d'ordre inférieur de façon à obtenir  $\lambda_{(2)}$  sous la forme  $\lambda_{(1)} + (X^* T_1^1 X^*)^{-1} (\tau_x - \frac{1}{2} \tau_x^{(1)})$  où  $T_1 = \text{diag}(h_k d_{(1)}^k, d_{(1)}^k) = (U - 1)^{-1}(1 - L)^{-1}(U - g_k^{(1)})(g_k^{(1)} - L)$ . D'autres itérations sont effectuées de façon semblable jusqu'à ce que les CE soient respectées. Le vecteur poids  $c_{GMDI}$  est proche de  $h$  du fait que, sous réserve des CE, l'itération ci-dessus correspond à l'algorithme de Newton-Raphson pour la minimisation de la fonction de distance  $\Delta_{GMDI}(c, h)$  donnée par  $A^{-1} \sum_{k=1}^n h_k [(g_k^{(1)} - L) \log\{(1 - L)^{-1}(g_k^{(1)} - L)\} + (U - g_k^{(1)}) \log\{(U - 1)^{-1}(U - g_k^{(1)})\}]$ . On trouvera à l'annexe A7 l'algorithme de calcul pour GMDI.

## 3. EXEMPLES NUMÉRIQUES

### 3.1 Description des données

Nous envisageons l'application des sept méthodes de correction décrites ci-dessus à des données tirées de l'Enquête de 1990 sur les dépenses des familles (FAMEX) de Statistique Canada pour deux villes (ou domaines), soit Regina et Saskatoon en Saskatchewan. Quatre variables sont à l'étude: les dépenses annuelles de réparation et de rénovation pour les logements en propriété, le mobilier et l'équipement, les vêtements pour femmes, les vêtements pour hommes. L'enquête FAMEX s'ajoute à l'Enquête sur la population active (EPA) du Canada et, par conséquent, elle se fonde sur le plan EPA – un échantillon en grappes de ménages stratifié à degrés multiples (voir Singh et coll. 1990). Les échantillons sont prélevés indépendamment dans deux villes, soit Regina et Saskatoon. Pour les deux villes, le nombre de strates est de 30 et 34 respectivement, et le nombre d'unités primaires d'échantillonnage (UPH) choisies dans

à se trouver exactement à la limite. Les paramètres modèles sont choisis par itération. Au début, nous posons  $\phi_k^{(0)} = 1$  et, pour l'itération 1,  $\lambda_{(1)} = \lambda_{MCS-n}$  de façon à obtenir  $g_k^{(1)} = (1 + \phi_k^{(0)} x_k^* \lambda_{(1)})$ , qui est corrigé d'avantage (ou tronqué) de façon à ce que l'on obtienne  $g_k^{(1)}$  sous la forme  $(1 + \phi_k^{(1)} x_k^* \lambda_{(1)})$  où  $\phi_k^{(1)} = \phi_k^{(0)} \phi_k^{(1)}$ , de sorte que les RAFV sont respectées. Toutefois, il est possible que  $g_{(1)}$  ne respecte pas les CE. À noter que la différence entre  $g_{(1)}$  et  $g_{MCS-n}$  est plus faible. Pour l'itération 2,  $\lambda_{(1)}$  est corrigé selon un terme d'ordre inférieur (uniformément pour  $k$ ) de façon à définir  $\lambda_{(2)}(k)$  sous la forme  $\lambda_{(1)} + (1/\phi_{(1)}^{(1)})(X^* T_1^1 X^*)^{-1} (\tau_x - \tau_x^{(1)})$ , où  $T_1 = \text{diag}(h)$  sauf que les éléments diagonaux sont tronqués à zéro pour tous les  $k$  pour lesquels  $\phi_k^{(1)} < 1$ , c'est-à-dire les unités qui ont été tronquées à l'itération précédente. Cette exclusion des éléments diagonaux ressemble quelque peu à l'utilisation d'un facteur de cadrage zéro pour la méthode SMCs. Pour la deuxième itération, nous avons  $g_k^{(2)} = 1 + \phi_k^{(1)} x_k^* \lambda_{(2)}(k)$ , et les facteurs de troncature  $\phi_k^{(2)}$  sont utilisés de façon à ce que l'on obtienne des  $g_k^{(2)}$  qui respectent les RAFV. Les itérations subséquentes se définissent de façon semblable. Il est clair que, contrairement à SM, les RAFV sont respectées ici à chaque itération. Les itérations se poursuivent jusqu'à ce que les CE soient respectées. Le vecteur poids,  $c_{MCS-r}$  est proche de  $h$  car les itérations définies ci-dessus représentent les étapes de Newton-Raphson pour la minimisation de la fonction de distance  $\Delta_{MCS-r}(c, h) = \sum_k (c_k - h_k)^2 / h_k$  si  $Lh_k \leq c_k \leq Uh_k$ ;  $\infty$  autrement, sous réserve des CE. On trouvera l'algorithme de calcul à l'annexe A5. À noter que, dans la pratique, il est plus commode d'utiliser  $g_k^{(v)}$  directement sans avoir à calculer  $\phi_k^{(v)}$  séparément.

## 2.6 MÉTHODE 6 (information de discrimination

modifiée restrictive ou MDI-r)

Cette méthode, qui relève également de la famille II, a été proposée par Singh (1993) suivant Deville et Sæmdal (1992) dans l'élaboration de MCS-r. Elle se rapporte à MDI-n de la même façon que MCS-r se rapporte à MCS-n. Il s'agit essentiellement de corriger les paramètres  $\phi$  et  $\lambda$  dans le coefficient de correction  $g_k = \phi_k \exp(x_k^* \lambda)$  de façon que les RAFV et les CE soient respectées. Le paramètre de troncature  $\phi$  ressemble à celui de MCS-r. Cela se fait par itération. Comme pour MCS-r, pour l'itération 1 nous posons  $g_{(1)}^{(1)} = \phi_{(0)}^{(1)} \exp(x_{(1)}^* \lambda_{(1)})$  où  $\phi_{(0)}^{(1)} = 1$ ,  $\lambda_{(1)} = \lambda_{MCS-n}$ , que l'on corrige d'avantage selon un terme d'ordre inférieur de façon à obtenir  $g_{(1)}^{(1)}$  sous la forme  $\phi_{(1)}^{(1)} \exp(x_{(1)}^* \lambda_{(1)})$  afin que les RAFV soient respectées, c'est-à-dire qu'il se trouve à l'intérieur de  $[L, U]$ . Ensuite, pour l'itération 2, on corrige  $g_{(1)}^{(1)}$  selon un terme d'ordre inférieur de façon à obtenir  $g_{(2)}^{(1)}$  sous la forme  $\phi_{(2)}^{(1)} \exp(x_{(1)}^* \lambda_{(2)})$ , où  $\lambda_{(2)} = \lambda_{(1)} + (X^* T_1^1 X^*)^{-1} (\tau_x - \frac{1}{2} \tau_x^{(1)})$ , et  $T_1 = \text{diag}(h_k g_k^{(1)})$ , sauf que les éléments diagonaux sont tronqués à zéro pour tous les  $k$  pour lesquels  $\phi_k^{(1)} < 1$ . Les facteurs de troncature  $\phi_k^{(2)}$  sont tronqués à zéro pour tous les  $k$  pour lesquels  $\phi_k^{(1)} < 1$ . Les facteurs de troncature  $\phi_k^{(2)}$  permettent de s'assurer que les RAFV sont respectées. Les itérations se poursuivent jusqu'à la convergence, c'est-à-dire jusqu'à ce que les CE soient respectées. Le vecteur poids  $c_{MDI-r}$  est proche de  $h$  car les itérations définies ci-dessus représentent les étapes de Newton-Raphson pour la



## 2.3 MÉTHODE 3 (Huang-Fuller modifiée ou chi carré modifiée mis à l'échelle, SMCS)

Cette méthode relève de la famille I (cas restrictif) et représente une légère modification de la méthode attribuée à Huang et Fuller (voir Singh 1993; voir aussi Fuller, Loughin et Baker 1994). Comme pour la régression, on suppose que le modèle du coefficient de correction est linéaire pour  $x$ . Afin que ces corrections satisfassent bien les RAFV, on utilise un facteur de cadrage  $q_k$  ( $0 < q_k \leq 1$ ) pour chaque  $k$  de façon à réduire le changement des poids  $h$  pour les unités dont les  $g_k$  ont tendance à sortir des limites  $[L, U]$ . Ainsi, le poids  $g$  est représenté par  $g_k = 1 + q_k x_k \lambda$  où les paramètres modèles  $q$  et  $\lambda$  sont choisis par itération, c'est-à-dire que l'on trouve  $\lambda$  pour un  $q$  donné pour ensuite trouver  $q$  pour un  $\lambda$  donné. Nous commençons par  $q_k^{(0)} = 1$  pour tous les  $k$  et nous posons  $\lambda_{(1)} = \lambda_{MCS-n}$  pour l'itération 1. Or il est clair que  $c^{(1)}$  satisfait aux CB, mais qu'il n'est pas nécessaire de satisfaire aux RAFV. Compte tenu de la position des  $g_k$  par rapport à  $[L, U]$ , on peut utiliser une règle empirique pour définir les  $q_k$  de façon que les  $q_k$  affectent davantage les unités qui sont plus éloignées des limites que celles qui sont plus rapprochées. Les facteurs de cadrage  $q_k^{(1)}$  ainsi déterminés définissent à leur tour  $\lambda_{(2)}$  pour l'itération 2 sous la forme  $(X^T X)^{-1} (x_k - \bar{x}_k) = \text{diag}(q_k^{(1)} h_k)$ ,  $q_k^{(1)} = q_k^{(0)} q_k^{(2)}$  respectant les CB après itération. À noter que, en vertu des conditions de régularité habituelles,  $\lambda_{(2)}$  diffère de  $\lambda_{(1)}$  uniquement selon un terme d'ordre inférieur, puisque la différence absolue maximale  $|q_k^{(1)} - 1|$  est faible. Ensuite, si  $c^{(2)}$  ne respecte pas les RAFV après l'itération 2, les facteurs de cadrage  $q_k^{(2)}$  sont définis convenablement et s'ajoutent à  $q_k^{(1)}$  pour donner  $q_k^{(2)}$  en vue de l'itération 3. On obtient ensuite le  $\lambda_{(3)}$  pour l'itération 3 comme auparavant de sorte que les CB sont respectées après l'itération. Les itérations se poursuivent jusqu'à ce qu'il y ait convergence, c'est-à-dire jusqu'à ce que les RAFV soient respectées. Le vecteur poids  $c_{SMCS}$  est proche de  $h$  car à chaque itération  $v \geq 1$ ,  $c^{(v)}$  minimise la fonction de distance  $\Delta_{SMCS}^v(c, h) = \sum_{k=1}^n (c_k - h_k)^2 / h_k q_k^{(v-1)}$  sous réserve des CB, où  $q_k^{(v-1)} = q_k^{(0)} q_k^{(1)} \dots q_k^{(v-1)}$  pour  $v \geq 1$ . À noter que, contrairement aux méthodes antérieures, la fonction de distance varie d'une itération à l'autre.

On trouvera à l'annexe A3 l'algorithme de calcul pour SMCS. À noter que, dans l'algorithme,  $[L, U]$  est rétrécie à  $[L', U']$  à l'aide d'un paramètre  $\alpha$  où  $L' = \alpha L + 1 - \alpha$ ,  $U' = \alpha U + 1 - \alpha$ , et  $0 < \alpha \leq 1$ . Cela suppose que l'on exclue aussi certaines unités qui se trouvent à l'intérieur de  $[L, U]$  mais proche de la limite. La convergence se trouve ainsi accélérée. On introduit également un autre paramètre  $\beta$ ,  $0 \leq \beta \leq 1$  afin de permettre une exclusion différentielle de diverses unités.

## 2.4 MÉTHODE 4 (rétrecissement-minimisation, SM)

Cette méthode bien connue relève de la famille II dans le

cas restrictif; elle provient de Deville et Särndal (1992). Comme pour la méthode SM, le modèle du coefficient de correction est supposé linéaire pour  $x$  avec un nouveau paramètre appelé facteur de troncature  $\phi_k$  ( $0 < \phi_k \leq 1$ ) utilisé pour chaque  $k$  de façon que les  $g_k$  respectent les RAFV, c'est-à-dire que l'on pose  $g_k$  à  $(1 + \phi_k x_k) \lambda(k)$ . La seule différence entre le facteur de troncature  $\phi_k$  utilisé ici et le facteur de rétrécissement utilisé pour SM est qu'ici les poids  $g$  qui se trouvent à l'extérieur de  $[L, U]$  sont toujours corrigés de façon

## 2.5 MÉTHODE 5 (tronqué linéaire ou chi carré modifiée restrictif, MCS-r)

On trouvera l'algorithme de calcul à l'annexe A4. Il ne faut pas oublier que, dans la méthode ci-dessus, lorsqu'un poids  $g$  se trouve à l'extérieur des limites  $L$  et  $U$ , une correction permet de ramener le poids  $g$  à la limite  $L$  ou  $U$ . L'introduction d'un nouveau paramètre  $\alpha$  ( $0 < \alpha \leq 1$ ) permet à l'utilisateur de ramener le poids  $g$  plus à l'intérieur de la limite jusqu'à un point  $L'$  ou  $U'$  ( $L' = \alpha L + 1 - \alpha$ ,  $U' = \alpha U + 1 - \alpha$ ). Cela ressemble quelque peu au paramètre  $\alpha$  de SMCS. L'introduction d'un autre paramètre  $\eta$  ( $0 < \eta \leq \alpha \leq 1$ ) permet de corriger les poids  $g$  au niveau de  $L'$  ou de  $U'$  également pour les unités qui se trouvent à l'intérieur de  $[L, U]$ , mais proche de la limite quand elles se trouvent à l'extérieur de  $[L'', U'']$  où  $L'' = \eta L + 1 - \eta$ ,  $U'' = \eta U + 1 - \eta$ . Tous ces paramètres aident à accélérer la convergence de façon générale.

distance dépend de l'itération.

remment. Comme pour la méthode SMCS la fonction de  $c^{(v)}$  directement de  $c^{(v)}$  sans avoir à calculer  $\psi^{(v)}$  séparément des CB. À noter que, dans la pratique, on peut obtenir réserve des CB,  $\Delta_{SM}^v(c, h) = \sum_{k=1}^n (c_k - h_k)^2 / c_k^{(v-1)}$  sous distance,  $\Delta_{SM}^v(c, h) = \sum_{k=1}^n (c_k - h_k)^2 / c_k^{(v-1)}$  minimise la fonction de chaque itération  $v \geq 1$ ,  $c^{(v)}$  minimise la fonction de soient respectées. Le vecteur poids  $c_{SM}$  se rapproche de  $h$  car jusqu'à la convergence, c'est-à-dire jusqu'à ce que les RAFV inférieure uniformément pour  $k$ . Les itérations se poursuivent après l'itération, mais pas nécessairement les RAFV. À noter que  $\lambda_{(2)}(k)$  diffère de  $\lambda_{(1)}$  selon un terme d'ordre inférieur utilisant les poids  $c^{(1)}$ . Encore une fois les CB sont respectés, et  $\frac{x_k}{c^{(1)}}(1 - \frac{x_k}{c^{(1)}})$  est l'estimateur avec facteur d'extension  $h_k g_k$ , et  $\frac{x_k}{c^{(1)}}(1 - \frac{x_k}{c^{(1)}})$  où  $\Gamma_1 = \text{diag}(c^{(1)} x_k)$ ,  $c_k^{(1)} = (X^T X)^{-1} (x_k - \bar{x}_k) = \text{diag}(q_k^{(1)} h_k)$ ,  $q_k^{(1)} = q_k^{(0)} q_k^{(2)}$  respectant les CB après itération. À noter que, en vertu des conditions de régularité habituelles,  $\lambda_{(2)}$  diffère de  $\lambda_{(1)}$  uniquement selon un terme d'ordre inférieur, puisque la différence absolue maximale  $|q_k^{(1)} - 1|$  est faible. Ensuite, si  $c^{(2)}$  ne respecte pas les RAFV après l'itération 2, les facteurs de cadrage  $q_k^{(2)}$  sont définis convenablement et s'ajoutent à  $q_k^{(1)}$  pour donner  $q_k^{(2)}$  en vue de l'itération 3. On obtient ensuite le  $\lambda_{(3)}$  pour l'itération 3 comme auparavant de sorte que les CB sont respectées après l'itération. Les itérations se poursuivent jusqu'à ce qu'il y ait convergence, c'est-à-dire jusqu'à ce que les RAFV soient respectées. Le vecteur poids  $c_{SMCS}$  est proche de  $h$  car à chaque itération  $v \geq 1$ ,  $c^{(v)}$  minimise la fonction de distance,  $\Delta_{SMCS}^v(c, h) = \sum_{k=1}^n (c_k - h_k)^2 / h_k q_k^{(v-1)}$  sous réserve des CB, où  $q_k^{(v-1)} = q_k^{(0)} q_k^{(1)} \dots q_k^{(v-1)}$  pour  $v \geq 1$ . À noter que, contrairement aux méthodes antérieures, la fonction de distance varie d'une itération à l'autre.

On trouvera à l'annexe A3 l'algorithme de calcul pour SMCS. À noter que, dans l'algorithme,  $[L, U]$  est rétrécie à  $[L', U']$  à l'aide d'un paramètre  $\alpha$  où  $L' = \alpha L + 1 - \alpha$ ,  $U' = \alpha U + 1 - \alpha$ , et  $0 < \alpha \leq 1$ . Cela suppose que l'on exclue aussi certaines unités qui se trouvent à l'intérieur de  $[L, U]$  mais proche de la limite. La convergence se trouve ainsi accélérée. On introduit également un autre paramètre  $\beta$ ,  $0 \leq \beta \leq 1$  afin de permettre une exclusion différentielle de diverses unités.



d'autres noms aux méthodes existantes afin de faciliter la compréhension du lien entre différentes méthodes. Les noms retenus se fondent sur les mesures de distance bien connues utilisées dans l'analyse des données de dénombrement.

Il est à noter que, puisque toutes les méthodes sont asymptotiquement équivalentes à la méthode de régression, il est possible d'évaluer la variance asymptotique de  $\hat{\tau}_j$  pour chaque méthode à l'aide de  $\sum_k (\pi_k - \pi_k \pi_j) \pi_{k1}^1 (e_k g_k)(e_l g_l)$ , comme dans Deville et Särndal (1992, équation 3.4) où  $\pi_k, \pi_{kl}$  représentent respectivement les probabilités d'inclusion d'ordre 1 et 2,  $e_k$  représentent les résidus d'échantillon  $y_k - \mathbf{B} \mathbf{x}_k$  avec  $\mathbf{B}' = (\mathbf{y}' \mathbf{T}_0 \mathbf{X})(\mathbf{X}' \mathbf{T}_0 \mathbf{X})^{-1}$ , et  $\mathbf{T}_0$  représente la matrice diagonale  $(h) n \times n$ .

## 2.1 MÉTHODE 1 (régression linéaire ou chi carré modifié non restrictif, MCS-u)

Cette méthode, la plus simple, donne l'estimateur de régression généralisé populaire de Särndal (1980). Ici, le modèle du coefficient de correction est supposé linéaire pour  $\mathbf{x}, \mathbf{c}, \mathbf{a}, \mathbf{d}$ , que  $g_k = 1 + \mathbf{x}_k' \lambda$ , pour un  $p$ -vecteur quelconque de paramètres modèles  $\lambda$  qui respecte les CB. Autrement dit,  $\sum_{k=1}^n h_k (1 + \mathbf{x}_k' \lambda) x_{kj} = \tau_{xj}$ , pour tous les  $j$ . On obtient sous la forme  $(\mathbf{X}' \mathbf{T}_0 \mathbf{X})^{-1} (\tau_x - \frac{\tau}{\tau(0)})$ . Les poids  $c$  restent proches des poids  $h$  du fait que le choix ci-dessus de poids  $g$  minimise la fonction de distance,  $\Delta_{MCS-u}(\mathbf{c}, \mathbf{h}) = \sum_{k=1}^n (c_k - h_k)^2 / h_k$  sous réserve des CB. À noter que les poids  $g$  risquent d'être négatifs pour certains  $k$ . Cela est peu souhaitable dans la pratique même si la simplicité de la méthode lui donne un certain attrait. On trouvera à l'annexe A l'algorithme de calcul pour MCS-u.

## 2.2 MÉTHODE 2 (itération du quotient ou information de discrimination modifiée non restrictive, MDI-u)

Cette méthode est également répandue. Ici, le modèle du coefficient de correction  $g_k$  prend la forme  $\exp(\mathbf{x}_k' \lambda)$ , de sorte qu'il est forcément non négatif. Contrairement à la méthode 1, le vecteur de paramètres modèles  $\lambda_{MDI-u}$  est obtenu par itération de façon à satisfaire aux CB. Les itérations peuvent commencer par  $\lambda_{MCS-u}$  de l'estimateur de régression généralisé, c'est-à-dire que pour l'itération 1, on pose  $\lambda_{(1)} = \lambda_{MCS-u}$ , ce qui suppose  $c_{(1)}^k = h_k \exp(\mathbf{x}_k' \lambda_{(1)})$ . Ces poids  $c$ , de façon générale, ne respectent pas les CB. Pour ce qui est de l'itération 2 de cette méthode, le  $\lambda_{(1)}$  est corrigé (à l'aide d'un terme d'ordre inférieur) de façon à définir  $\lambda_{(2)}$  sous la forme  $\lambda_{(1)} + (\mathbf{X}' \mathbf{T}_1 \mathbf{X})^{-1} (\tau_x - \frac{\tau}{\tau(1)})$ , où  $\mathbf{T}_1 = \text{diag}(c_{(1)}^k)$ . Le terme  $\lambda$  se définit de façon semblable pour d'autres itérations jusqu'à la convergence, c'est-à-dire jusqu'à ce que les CB soient respectées. Les poids  $c$  restent proches des poids  $h$  puisque les itérations utilisées dans la méthode ci-dessus représentent les étapes de Newton-Raphson pour la minimisation de la fonction de distance,  $\Delta_{MDI-u}(\mathbf{c}, \mathbf{h}) = \sum_{k=1}^n [c_k \log(c_k/h_k) - c_k + h_k]$  sous réserve des CB. À noter que, même si les poids  $g$  sont non négatifs, ils peuvent être très élevés, ce qui dans la pratique est nettement peu souhaitable. On trouvera à l'annexe A2 l'algorithme de calcul pour MDI-u.

de comparer différentes méthodes de calage en fonction d'un ensemble de données réelles. Plus particulièrement, nous examinerons à l'aide d'une analyse descriptive l'effet des RAFV sur le nombre de calculs à effectuer, la distribution des coefficients de correction des poids, les estimations ponctuelles et leur variance. Les études comparatives connexes menées sur les méthodes de calage qui se fondent sur des ensembles de données réelles relèvent de Deville, Särndal et Santory (1993) et de Stukel et Boyer (1993). Toutefois, ces études se limitent à des méthodes de la famille II et elles portent principalement sur la distribution des coefficients de correction des poids. Enfin, on trouvera une discussion à la section 4.

## 2. JUSTIFICATION HEURISTIQUE DES ESTIMATEURS DE CALAGE

Soit  $n, N$  la taille de l'échantillon et la taille de la population respectivement. Nous notons  $h_k$  le poids initial ou le poids  $h$  (utilisé dans l'estimateur avec facteur d'extension ou de Horvitz-Thompson  $\sum_{k=1}^n y_k h_k$ ) pour le  $k$ -ième élément où  $y_k$  est la valeur de la variable à l'étude. On suppose que les poids  $h$  incorporent des corrections pour toute non-réponse. Le paramètre qui nous intéresse est la population totale pour  $y$ , notée par  $\tau_y$ . Pour chaque  $k$ , il existe des  $p$ -vecteurs de variables auxiliaires,  $\mathbf{x}_{kj}, j = 1, \dots, p$  pour lesquelles la population totale ou la contrainte d'échantonnage,  $\tau_{xj} = \sum_{k=1}^N x_{kj}$  pour chaque  $j$  est supposée connue. Le  $p$ -vecteur transposé  $\mathbf{x}_k$  indique  $(x_{k1}, \dots, x_{kp})$ , la  $k$ -ième ligne de la matrice  $n \times p, \mathbf{X}$ . Soit  $c_k^v$  le poids calé ou poids  $c$  pour le  $k$ -ième élément de la  $v$ -ième itération. À  $v = 0$ ,  $c_k^{(0)} = h_k$ . Les estimateurs avec facteur d'extension aux chiffres de population pour les variables  $y$  et  $\mathbf{x}_j$  avec les poids  $c$  à la  $v$ -ième itération sont notés  $\hat{\tau}_y^{(v)}$  et  $\hat{\tau}_{xj}^{(v)}$  respectivement.

Les RAFV sont précisées par la contrainte  $L \leq g_k \leq U$  où  $g_k = c_k/h_k$  et  $L < 1 < U$ , où  $L$  et  $U$  représentent des limites inférieure et supérieure convenables. Les coefficients de correction ( $p$ , ex. les  $g_k$ ) s'appellent également poids  $g$ . Nous considérons d'abord le cas non restrictif ( $p$ , ex. le calage sans RFAV) et, ensuite, le cas restrictif. Toutes les méthodes appliquées au cas restrictif exigent l'itération pour la solution. Nous supposons qu'il y a convergence après un nombre fini d'itérations.

Le critère de convergence se définit comme suit. Pour que l'itération respecte les RFAV, nous définissons un niveau de tolérance  $\epsilon$  ( $p$ , ex. 0.005 ou 0.01) pour la famille I tel que l'itération prend fin lorsque l'erreur relative absolue (BRA) maximale pour les RFAV est  $\leq \epsilon$ . De même, nous définissons un niveau de tolérance ( $\delta > 0$ ) pour la famille II afin de respecter les CB par itération. En effet, notre objectif principal n'est pas de minimiser la fonction de distance, mais bien de trouver une solution qui respecte les CB et les RFAV. En plus de  $\epsilon$  et de  $\delta$ , nous définissons un paramètre  $v_{\max}$  qui limite le nombre d'itérations.

Le présent article décrit sept méthodes, deux pour le cas non restrictif, deux pour le cas restrictif de la famille I et trois autres pour le cas restrictif de la famille II. Nous avons donné



# Comprendre les estimateurs de calage dans les enquêtes par échantillonnage

A.C. SINGH et C.A. MOHL<sup>1</sup>

## RÉSUMÉ

Il existe des méthodes bien connues dues à Deville et Särndal (1992) qui permettent de corriger les poids d'échantillonnage de façon à respecter les contraintes d'étalement de calage. Il existe également une méthode antérieure, peut-être moins bien connue, due à Huang et Fuller (1978). De plus, d'autres méthodes ont été élaborées par Singh (1993), qui a montré que, comme les résultats de Deville-Särndal l'indiquent, toutes ces méthodes sont asymptotiquement équivalentes à la méthode de régression. Le présent article comporte trois objectifs: i) fournir une justification heuristique simple pour tous les estimateurs de calage, y compris ceux qui sont bien connus et moins bien connus, en adoptant une stratégie non traditionnelle; ii) s'agit d'abord de choisir un modèle (au lieu de la fonction de distance) pour le coefficient de correction des poids et, ensuite, de montrer qu'une méthode appropriée d'ajustement des modèles correspond à la solution de minimisation de la distance; iii) offrir aux praticiens des algorithmes de calcul rapide; iv) comparer différentes méthodes du point de vue de la distribution des coefficients de correction des poids, des estimations ponctuelles, de la précision estimée et du nombre de calculs à effectuer en citant des exemples numériques fondés sur un ensemble de données réelles. Il est possible de formuler des observations intéressantes à l'aide d'une analyse descriptive de résultats numériques indiquant que, même si toutes les méthodes de calage ressemblent à la méthode de régression en ce qui concerne les limites larges, elles comportent des différences pour ce qui est des limites étroites.

**MOTS CLÉS:** Contraintes d'étalement; minimisation de la distance; poids non négatifs; restrictions applicables à la fourchette de valeurs.

## 1. INTRODUCTION

Lorsqu'on établit des estimations fondées sur des en-  
quêtes par échantillonnage, on corrige habituellement les  
poids d'échantillonnage afin d'obtenir des poids calés corres-  
pondant à des totaux ou à des contraintes d'étalement  
(CE) pour les variables auxiliaires. On utilise souvent la  
méthode de régression et la méthode itérative du quotient à  
cette fin. Même si ces méthodes ont de bonnes propriétés  
asymptotiques (voir Deville et Särndal 1992), elles peuvent  
donner des poids calés dont les propriétés (échantillon fini)  
sont indésirables. La méthode de régression peut donner des  
poids négatifs, tandis que la méthode itérative du quotient  
risque de produire des poids très élevés. C'est pourquoi, en  
présence de poids calés, on impose parfois des restrictions  
applicables à la fourchette de valeurs (RAFV). Il serait utile  
de disposer d'une méthode de calage qui i) produise des poids  
calés qui se rapprochent des poids d'échantillonnage ori-  
ginaux, grâce par exemple à la minimisation d'une fonction  
de distance appropriée entre les deux ensembles de poids, qui  
ii) respecte les CE et qui iii) répond aux RAFV. Il existe  
plusieurs méthodes documentées de correction des poids dans  
le cadre des CE et des RAFV (voir par exemple, Deville et  
Särndal (1992) au sujet des progrès récents, et Huang et  
Fuller (1978) au sujet des progrès antérieurs). On trouvera  
une synthèse, ainsi que des travaux plus poussés, dans Singh  
(1993). Il s'agit là de méthodes d'itération que l'on peut

classer en deux familles. La famille I comporte des méthodes  
qui respectent les CE après chaque itération, l'itération se  
poursuivant jusqu'à ce que les RAFV soient respectées. La  
famille II, pour sa part, comporte des méthodes qui respectent  
les RAFV après chaque itération, l'itération se poursuivant  
jusqu'à ce que les CE soient respectées. Les méthodes de  
Deville et Särndal (1992) relèvent de la famille II, tandis que  
la méthode de Huang et Fuller relève de la famille I. Singh  
(1993) a proposé deux autres méthodes, une pour chaque  
famille. À l'aide d'arguments semblables à ceux de Deville et  
Särndal (1992), Singh a élargi le résultat remarquable de  
Deville et Särndal en montrant que toutes les méthodes des  
familles I et II sont asymptotiquement équivalentes à la  
méthode de régression.  
À la section 2, une stratégie non traditionnelle permet de  
présenter chaque méthode qui pourrait aider à comprendre les  
estimateurs de calage. Une présentation heuristique de l'aspect  
fonctionnel du coefficient de correction des poids est suivie  
de l'établissement d'un lien entre une méthode appropriée  
d'ajustement des modèles et la minimisation de la fonction  
de distance. Du même coup, des algorithmes de calcul rapide  
sont mis à la disposition des praticiens. On pourra obtenir  
du deuxième auteur un programme informatique en logiciel  
GAUSS (voir également Singh et Mohl 1997). À la section 3,  
des exemples numériques permettent d'illustrer différentes  
méthodes à l'aide de données de l'Enquête sur les dépenses  
des familles (FAMEX) de Statistique Canada. Il est intéressant

Eltinge et Jang émettent des suggestions sur la façon d'évaluer la stabilité des estimations des composantes de la variance (plus précisément, les estimateurs de la variance intra-UPF) et d'autres valeurs apparentées dans un plan d'échantillonnage complexe à trois phases. Les auteurs envisagent pour mesures, un simple estimateur de la variance reposant sur le plan de sondage de l'estimateur de la variance intra-UPF et une approche basée sur l'estimation du nombre de «degrés de liberté». Une technique de simulation permet d'établir si la stabilité observée est cohérente avec les hypothèses habituelles concernant la stabilité de l'estimateur de la variance. Les auteurs appliquent les méthodes proposées aux données de la NHANES III et démontrent que les propriétés de la stabilité véritable peuvent varier sensiblement d'une variable à l'autre, et que les estimateurs de la variance intra-UPF peuvent être beaucoup moins stables que le simple dénombrement des unités secondaires de chaque strate le laisse supposer.

Berger discute du plan de Chao qui permet de sélectionner de façon séquentielle, un échantillon sans remise à probabilités inégales et à taille fixe. Dans ce contexte, il suggère une approximation des probabilités d'inclusion d'ordre deux en vue d'obtenir un estimateur de la variance approché pour l'estimateur de Horvitz et Thompson. Cette variance est ensuite comparée à des approximations données pour d'autres procédures ou plan de sélection. Des conditions d'équivalence de ces approximations sont présentées.

Cowling, Chambers, Lindsay et Parameswaran présentent deux techniques servant à produire des données lissées spatialement. Ils en examinent les implications au niveau de l'estimation de petites et de grandes régions. Dans le premier cas, on lisse les poids de l'échantillon spatialement au moyen d'une méthode de régression linéaire modifiée qui réduit la variance mais augmente le biais des estimations. Pour les grandes régions, on recourt à une méthode de régression non paramétrique afin de lisser spatialement les données qui servent ensuite à produire une carte au moyen d'un système d'information géographique. Les auteurs donnent les résultats d'une simulation grâce à laquelle on a cherché à établir la méthode et le niveau de lissage qui conviennent le mieux aux cartes.

Brick, Wakseberg et Keeler proposent d'utiliser les renseignements sur les interruptions du service téléphonique de manière à ajuster les estimations d'enquête pour compenser le biais de sous-dénombrement. Les données recueillies sur les interruptions de service téléphonique permettent de réduire le biais, mais parallèlement la variance est susceptible d'augmenter en raison d'une plus grande variabilité des poids de sondage. Les résultats obtenus d'une enquête à l'échelle nationale montrent un potentiel non négligeable de diminution de l'erreur quadratique moyenne des estimations, sous certaines conditions.

Enfin, Pandher recourt à une approche de modélisation pour déterminer le fractionnement optimal d'une population enquêtée en strates à tirage sans remise et à tirage partiel. L'approche suppose qu'on ne s'intéresse qu'à une variable et qu'on recourt à une méthode d'échantillonnage à probabilité proportionnelle à la taille pour les strates à tirage partiel. L'auteur propose un algorithme pour déterminer le point de séparation optimal entre les groupes à tirage sans remise et à tirage partiel. Une exigence primordiale de l'algorithme est que l'espérance de la variance du modèle soit une fonction convexe du nombre d'unités dans les strates à tirage sans remise, lequel repose sur les hypothèses à la base du modèle et la forme des probabilités d'inclusion. La méthode est ensuite appliquée à l'Enquête sur les finances des administrations locales de Statistique Canada.

Le rédacteur en chef



## Dans ce numéro

Ce numéro de *Techniques d'enquête* débute par une partie spéciale réunissant quatre articles sur la **pondération et l'estimation**.

Le premier article, rédigé par Singh et Mohl, donne un aperçu des méthodes de calage sous un angle inhabituel, l'objectif étant de parvenir à une meilleure compréhension heuristique des méthodes en question. Pour Deville et Särndal, les méthodes de calage servent à réduire au minimum l'écart global entre les poids finaux et les poids d'échantillonnage, sous réserve que l'estimation du total de certaines covariables puisse être séparée à des taux de population connus. Singh et Mohl proposent d'autres méthodes de calage, qu'ils dérivent de différents modèles pour les facteurs de correction des poids. Les algorithmes de calcul de diverses méthodes sont présentés en annexe et un exemple numérique montre comment les facteurs de correction des poids résultants peuvent varier d'une méthode à l'autre.

Stukel, Hidiroglou et Särndal s'intéressent aussi aux estimateurs de calage, la classe des estimateurs ponctuels basés sur le plan de sondage mis au point par Deville et Särndal. Ces estimateurs, dérivés des fonctions de distance, permettent de restreindre les poids finaux afin de leur donner une valeur positive ou de les plafonner, ce qui évite le problème habituel des poids négatifs qui survient lorsqu'on recourt à un estimateur de régression. À l'aide de simulations, les auteurs examinent les propriétés de plusieurs estimateurs de ce genre reposant sur diverses fonctions de distance; ils insistent en particulier sur les propriétés des estimateurs de la variance correspondants, notamment l'estimateur jackknife et l'estimateur de Taylor. Il est surprenant de constater que les estimateurs ponctuels et les estimateurs de la variance correspondants présentent un biais minime, même lorsqu'on restreint considérablement les poids finaux.

Jayasuriya et Valliant comparent trois façons de dériver les poids du ménage de l'Enquête sur les dépenses des consommateurs du U.S. Bureau of Labor Statistics. Habituellement, les poids d'enquête sont calés en fonction des totaux démographiques pour diverses caractéristiques au niveau individuel, si bien que les membres d'un ménage reçoivent des poids finaux différents. En vertu de la méthode de la personne principale, on choisit le poids d'un membre du ménage comme poids final du ménage. Ensuite, on ajuste le vecteur des variables auxiliaires de chaque membre du ménage par régression, d'après la moyenne du ménage, de sorte qu'on obtient des poids calés identiques pour toutes les personnes du ménage. Une autre possibilité consiste à restreindre les facteurs de correction des poids afin d'éviter les valeurs extrêmes ou négatives. Les auteurs comparent les variantes de ces méthodes en fonction des poids finaux et des coefficients de variation estimés de plusieurs catégories de dépenses des ménages.

Dans le dernier article de la partie sur la **pondération et l'estimation**, Chen et Chen étudient le problème de l'estimation de l'intervalle de confiance pour la moyenne d'une population finie lorsqu'on possède des données auxiliaires. Rappelant les premiers résultats obtenus par Royall et Cumberland, selon lesquels un usage naïf des méthodes existantes articulées sur le plan de sondage débouche sur des intervalles de confiance caractérisés par de très piètres probabilités de couverture conditionnelle, les auteurs proposent des transformations qui permettent aux données de mieux respecter l'hypothèse de normalité sous-jacente, donc qui améliorent le taux de couverture. Les données auxiliaires sont intégrées de deux façons : soit directement, à l'inférence, quand les données sont connues pour chaque unité, soit par calage, selon la probabilité empirique, lorsque les données complémentaires ne sont connues qu'au niveau de la population. Les auteurs démontrent le bon fonctionnement de leurs méthodes par l'application d'une simulation à six populations réelles.

Dans leur article, Thompson et Fisher adaptent le test de McNemar à simple et à double échantillon afin qu'il s'applique aux données d'une enquête complexe. Ils se servent ensuite de leur version modifiée du test à double échantillon pour vérifier si le passage à une interview téléphonique assistée par ordinateur reposant sur un questionnaire remanié modifiera ou non l'estimation du chômage, d'après les données issues de l'étude à panel partiel du Current Population Survey du U.S. Bureau of the Census. Les auteurs parlent des résultats du test et les comparent aux résultats d'autres recherches concernant l'incidence de l'ITAO sur l'estimation du chômage.





# TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Volume 22, numéro 2, décembre 1996

## TABLE DES MATIÈRES

Dans ce numéro .....	105
<b>Pondération et estimation</b>	
A.C. SINGH et C.A. MOHL .....	107
Comprendre les estimateurs de calage dans les enquêtes par échantillonnage .....	107
D.M. STUKEL, M.A. HIDIROGLOU et C.-E. SÄRNDALE .....	117
Estimation de la variance des estimateurs de calage: comparaison des méthodes du jackknife et de la linéarisation de Taylor .....	117
B.R. JAYASURIYA et R. VALLANT .....	127
Application de l'estimation par régression restreinte dans une enquête-ménage .....	127
G. CHEN et J. CHEN .....	139
Une méthode de transformation applicable à l'échantillonnage de populations finies calée par une méthode de vraisemblance empirique .....	139
<hr/>	
K.J. THOMPSON et R. FISHER .....	149
Application des tests de McNemar à l'étude du panel partiel du Current Population Survey .....	149
J.L. ELTINGE et D.S. JANG .....	159
Mesures de la stabilité des estimateurs des composantes de la variance dans un plan d'échantillonnage stratifié à plusieurs degrés .....	159
Y.G. BERGER .....	169
Variance asymptotique pour un plan séquentiel sans remise à probabilités inégales .....	169
A. COWLING, R. CHAMBERS, R. LINDSAY et B. PARAMESWARAN .....	177
Applications du lissage spatial aux données d'enquête .....	177
J.M. BRICK, J. WAKSBERG et S. KEETER .....	187
Utilisation des données sur les interruptions du service téléphonique pour ajuster la couverture .....	187
G.S. PANDHER .....	201
Remaniement optimal du plan d'échantillonnage pour une population à distribution asymétrique au moyen d'un estimateur de régression généralisé avec applications .....	201
Remerciements .....	209

# TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Techniques d'enquête est répertoriée dans The Survey Statistician et Statistical Theory and Methods Abstracts. On peut en trouver les références dans Current Index to Statistics, et Journal Contents in Qualitative Methods.

## COMITÉ DE DIRECTION

### Président

G. J. Brackstone

### Membres

D. Binder  
G. J. C. Hole  
F. Mayda (Directeur de la Production)  
C. Patrick  
R. Platek (Ancien président)  
D. Roy  
M. P. Singh

## COMITÉ DE RÉDACTION

### Rédacteur en chef

M. P. Singh, *Statistique Canada*

### Rédacteurs associés

D. R. Bellhouse, *University of Western Ontario*  
D. Binder, *Statistique Canada*  
J.-C. Deville, *INSEE*  
J. D. Drew, *Statistique Canada*  
W. A. Fuller, *Iowa State University*  
R. M. Groves, *University of Maryland*  
M. A. Hidiroglou, *Statistique Canada*  
D. Holt, *Central Statistical Office, U.K.*  
G. Kalton, *Westat, Inc.*  
R. Lachapelle, *Statistique Canada*  
S. Linacre, *Australian Bureau of Statistics*  
D. Pfeffermann, *Hebrew University*  
J. N. K. Rao, *Carleton University*

### Rédacteurs adjoints

L. P. Rivest, *Université Laval*  
I. Sande, *Bell Communications Research, U.S.A.*  
F. J. Scheuren, *George Washington University*  
J. Sedransk, *Case Western Reserve University*  
R. Sitter, *Simon Fraser University*  
C. J. Skinner, *University of Southampton*  
R. Valliant, *U.S. Bureau of Labor Statistics*  
V. K. Verma, *University of Essex*  
P. J. Waite, *U.S. Bureau of the Census*  
J. Waksberg, *Westat, Inc.*  
K. M. Woller, *National Opinion Research Center*  
A. Zaslavsky, *Harvard University*

## POLITIQUE DE RÉDACTION

Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

## Présentation de textes pour la revue

Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à faire parvenir le texte rédigé en anglais ou en français au rédacteur en chef, M. P. Singh, Division des méthodes d'enquêtes des ménages, Statistique Canada, Tunney's Pasture, Ottawa (Ontario), Canada K1A 0T6. Prière d'envoyer quatre exemplaires dactylographiés selon les directives présentées dans la revue. Ces exemplaires ne seront pas retournés à l'auteur.

## Abonnement

Le prix de Techniques d'enquête (n° 12-001-XPB au catalogue) est de 45 \$ par année au Canada, 50 \$ (É.-U.) aux États-Unis, et de 55 \$ (É.-U.) par année à l'étranger. Prière de faire parvenir votre demande d'abonnement à Statistique Canada, Division des opérations et de l'intégration, Gestion de la circulation, 120, avenue Parkdale, Ottawa (Ontario), Canada K1A 0T6. Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête, l'American Association for Public Opinion Research et la Société Statistique du Canada.



# TECHNIQUES D'ENQUÊTE



## UNE REVUE ÉDITÉE PAR STATISTIQUE CANADA

DÉCEMBRE 1996 • VOLUME 22 • NUMÉRO 2

Publication autorisée par le ministre  
responsable de Statistique Canada  
© Ministre de l'Industrie, 1996

Tous droits réservés. Il est interdit de reproduire ou de transmettre  
le contenu de la présente publication, sous quelque forme ou  
par quelque moyen que ce soit, enregistré ou non, sur support  
magnétique, reproduction électronique, mécanique, photographique,  
ou autre, ou de l'emmagasiner dans un système de recouvrement,  
sans l'autorisation écrite préalable des Services de concession  
des droits de licence, Division du marketing,  
Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

Décembre 1996

Prix : Canada : 45 \$

États-Unis : 50 \$ US

Autres pays : 55 \$ US

N° 12-001-XPB au catalogue

Périodicité: semestrielle

ISSN 0714-0045

Ottawa









NUMÉRO 2

VOLUME 22

DÉCEMBRE 1996

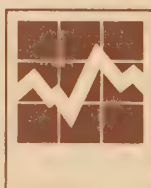
UNE REVUE  
ÉDITÉE  
PAR STATISTIQUE CANADA

Catalogue 12-001-X96

---

# TECHNIQUES D'ENQUÊTE

---



1 5 0 8











